PROTEIN EVOLUTION : THE DECIPHERING OF LATENT FACETS

by

DINABANDHU KUNDU

DEPARTMENT OF CHEMISTRY
INDIAN INSTITUTE OF TECHNOLOGY KANPUR
JUNE, 1994

PROTEIN EVOLUTION :
THE DECIPHERING OF LATENT FACETS

A Thesis Submitted
in Partial Fulfilment of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

by
DINABANDHU KUNDU

to the

DEPARTMENT OF CHEMISTRY
INDIAN INSTITUTE OF TECHNOLOGY, KANPUR

JUNE, 1994

DEDICATED

TO

MY PARENTS

STATEMENT

I hereby declare that the matter embodied in this thesis is the result of investigations carried out by me in the Department of Chemistry, Indian Institute of Technology, Kanpur, India, under the supervision of Professor S. Ranganathan.

In keeping with the general practice of reporting scientific observations, due acknowledgements have been made wherever the work embodied is based on the findings of other investigators.

IIT Kanpur

June, 1994

(Dinabandhu Kundu)

CERTIFICATE

Certified that the work contained in this thesis, entitled, "PROTEIN EVOLUTION: THE DECIPHERING OF LATENT FACETS" has been carried out by Mr. Dinabandhu Kundu, under my supervision and the same has not been submitted elsewhere for a degree.

S. Ranganathan
(S. Ranganathan)
Thesis Supervisor

IIT Kanpur
June, 1994

DEPARTMENT OF CHEMISTRY

INDIAN INSTITUTE OF TECHNOLOGY KANPUR, INDIA

CERTIFICATE OF COURSE WORK

This is to certify that Mr. Dinabandhu Kundu (Roll number 8920762) has satisfactorily completed all the courses required for the Ph.D. degree programme. These courses include:

CHM 605	Principles of Organic Chemistry
CHM 625	Principles of Physical Chemistry
CHM 645	Principles of Inorganic Chemistry
CHM 624	Modern Physical Methods in Chemistry
CHM 641	Advanced Inorganic Chemistry
CHM 681	Basic Biological Chemistry
CHM 800	General Seminar
CHM 801	Special Seminar
CHM 900	Post Graduate Research

Mr. Dinabandhu Kundu has successfully completed his Ph.D. qualifying examination on January, 1991, he also successfully presented his open seminar of the work embodied in this thesis.

P. K. Ghosh

Dr. P. K. Ghosh
Professor and Head
Department of Chemistry
I.I.T.Kanpur

S. Sarkar

Dr. S. Sarkar
Professor and Convener
Departmental Post Graduate Committee
Department of Chemistry
I.I.T.Kanpur

ACKNOWLEDGEMENTS

Paradoxical!! This section though appearing at the beginning of the dissertation is always written at the last. However, in retrospect, it gives me great pleasure to be able to show my appreciations and gratitude, in a very small way, to all those who have helped me and inspired me over the years.

At the very outset, I wish to express my heartfelt gratitude to my supervisor Professor S. Ranganathan for his encouragement, invaluable suggestions, dynamic and creative guidance which in a way revived my interest in the subject. I shall be forever indebted to him for introducing me to the challenging and enchanting world of Bioorganic Chemistry. His never ending stream of creative ideas and critical insights helped me to consolidate my grasp over both theoretical and practical aspects of Chemistry and Molecular Biology. This dissertation is an outcome of his exemplary interest and contagious enthusiasm. I consider it to be my good fortune to be under his tutelage. His potential to work constantly and systematically round the clock will always be a source of inspiration to me. Having said so far, I still say I do not have enough words to thank him.

I thank Dr.(Mrs) D. Ranganathan for her perspicuous comments and invaluable suggestions during the entire course of my Ph.D. work. I shall always cherish memories of her cogent advice, personal care and enthusiastic support.

A special thanks to all my colleagues in the laboratory Dr. G. P. Singh, Dr. Sujata Saini, Dr. Kavita Shah, Jayaraman, Bhisma, Tamilarasu, Shaji and Narendra for their pleasant association and various help which they rendered to me and also for keeping the laboratory atmosphere cheerful and conducive to good research work. It is indeed a privilege to thank Mr. Anand Ranganathan for his cheerful company during the early stage of the work.

I am grateful to Prof. D. Balasubramanian and Dr. R. Nagaraj, CCMB, Hyderabad for providing me an opportunity to work at CCMB and for their valuable discussions. I

am also grateful to Mr. V. M. Dhople for experimental help, HPLC separation, amino acid analysis and for his cheerful company during my stay at CCMB, Hyderabad.

It is my privilege to be able to express my gratitude and respect to Mr. S. Biswas, Education Department, Govt. of West Bengal, Prof. R. N. Mukherjee, Department of Chemistry, IIT Kanpur, Prof. P. K. Sen, Presidency College, Calcutta and Prof. P. Chakravorty, Scottish Church College, Calcutta. I thank them for their afflatus and keen interest in my academic progress.

I thank all the faculty members of the Department of Chemistry, IIT Kanpur, especially Prof. J. Iqbal and Prof. P. K. Bharadwaj for their advice and encouragement.

I shall be failing in my duty if I do not thank Samiran, Shanti, Sujay, Tapan and Debnath as they were a source of constant encouragement and inspiration to me. They shared my sorrows and joys making sure I do not lose track - a very special thanks to them. I would like to thank Manabendrada, Sujay, Shanti and Bhaskar for the computational help they rendered to me.

It is also my good fortune to be able to thank friends like Debashis, Tapas, Pulakesh, Pinaki, Kaustav, Debraj, Viswanath, Sankarshan, Chandreyi and Debalina who went out of their ways to help me and were pleasant company. Life here, without them would have been insipid and dull.

It is my privilege to thank Sangita and Romi; they have made my stay in IITK refreshing, enchanting, stimulating and memorable one.

I take this opportunity to thank S.K.Pandey, Govindaraju, Immanuel, Illangovan, Kashi, Arun, Promad, Subratada, Mama, Indrani, Tapan Khan, Subit, Nilu, Punniyamurthy, Manoj, Sanjoy, Kalyanraman, Susanta, Asif, Samarda, Tapan Lal, Debashis Bandyopadhyay, Tapobrato, Prasenjit and all others who have helped me in some way or other during my stay here.

I would like to express my thanks to Mrs. Mita Mukherjee for making my stay here as homely as possible.

My thanks are due to my friends Asish and Goutam, IACS, Calcutta, Partha, IISc, Bangalore, and Gangadhar, IPCL for their constant support in my all endeavours.

I shall forever be thankful to "Bodhi" - a Bengali association at IITK, which has a Bengali library and organizes cultural programmes, for making my stay out here enjoyable and interesting.

I also express my sincere thanks to Santosh and Panditji for the endless cups of tea, coffee and all other extra facilities that they extended to me.

I would like to thank all the staff members of the Chemistry Department for their cordial assistance. I extend my thanks to Mr. B. K. Jain for his beautiful art work presented in this thesis, Mr. A. Bhavsor and Umesh for recording NMR, Mr. N. Ahmad for IR and elemental analysis, to the personnel of glass blowing and liquid nitrogen plant for their ready help and RSIC unit of CDRI, Lucknow for FAB mass spectra and elemental analysis.

Coming to the concluding portion of this acknowledgement, I would like to mention one person, without her this acknowledgement will remain ever incomplete - thank you Indira.

Finally, I thank my parents, sisters and brother for what I am and for accepting me for what I am not. Trying to thank them in a few words would serve nothing but to disparage their contributions in making me what I am today. This is a debt, I feel proud to remain indebted with.

CONTENTS

Abbreviation

1. Code bases are represented by the standard one letter code.

2. Representation of amino acids and their derivatives:

a. All amino acids are represented by standard three letter code and one letter code. eg: Pro-Leu (P-L) - represents a peptide formed from the amino acid proline and leucine.

b. A symbol to the left and hyphenated is a blocking group on the alpha amino group. eg: Ts-Pro = N^{α}-Tosyl-proline.

c. A symbol to the right and hyphenated is a C-terminal protection. eg: Pro-OMe = methyl ester of proline and Pro-OH represents simple proline.

Other abbreviations used in this thesis are as follows:

Ac_2O	Acetic anhydride
Boc	tert-butyloxy carbonyl
DCC	dicyclohexyl carbodiimide
DCU	dicyclohexyl urea
DMF	dimethyl formamide
DNA	deoxyribonucleic acid
EtOAc	ethyl acetate
Et_3N	triethylamine
FAB	fast atom bombardment
g	gram
h	hour
HOBt	hydroxybenzotriazole
HPLC	high pressure liquid chromatography
ir	infra red
MeOH	methanol
min	minute
mL	mililitre

mmol	milimole
mp	melting point
ms	mass spectroscopy
nm	nanometer
nmr	nuclear magnetic resonance
OMe	methyl ester
ppm	parts per million
py	pyridine
rt	room temperature
Ts	p-toluene sulfonyl
TsCl	p-toluene sulfonyl chloride
tlc	thin layer chromatography
TMS	tetramethylsilane
WSCDI	water soluble carbodiimide, 1, Cyclohexyl-3- (2-morpholinoethyl) carbodiimide metho-p-toluenesulfonate

A. INTRODUCTION

In hierarchical terms the epitome of the manifestations of evolution transcending millennia is, most assuredly the symbiotic relationship between the informational and the functional system, which is the foundation for all living organisms as attested by the facet that the protocols involved here necessarily calls for a cascade of events which, in cadence, has to be tuned with precision and with practically no room what so ever for false notes.

The tracking of the very roots leading to such manifestation, whilst imperative and of great potential utility, is beset with, not only myriads of problems, but also the certainty that the real truth will never be clearly known, since evolution by necessity obliterates the past.

In spite of such bleak and arid scenario, valiant assaults have been made on this bastion of latent knowledge. The endeavours presented in the thesis form part of one such expedition. The approach here is rooted in the domain of organic chemistry and the belief that the principles pertaining to interactions in this area would have a direct bearing on the problem at hand. In specific terms the evolution of the functional system and the genesis of the information function composite forms the focal theme, since, these events are surely harbingers and landmarks in the long and intricate passage leading to the primitive cell, the fundamental unit of all living organisms.

The procedure adopted here is analytical and logical comprising of a model for possible interactions predating events preceding the evolution of the genetic code, a landmark, that is principle froze the pattern of life.

The search of imprints of protein evolution and that of the information-function composite, as latent facets constitutes the cornerstone of investigation reported here.

Using experiments and theory, deductive logic has been brought to the fore to provide a rationale that would often a cogent and cohesive explanation.

The mural presented here encompasses a vast stretch starting from the delineation of intrinsic preference for peptide bond formation, involving the coded amino acids to an analysis of zinc finger modules which even presently are involved in information-function recognition.

Surely, the mechanisms for the peptide bond formation are markers of evolution. In the post genetic code era the protocols governing the ribosomal protein synthesis orchestrate a rather unusual chemical creation. An event so complicated and so precise must have evolved and, ironically, for logical reasons the processes here are some what at variance with principles of organic synthesis that vie for optimum energy utilization. The latter principles may well have governed the early stages of protein evolution. Indeed the remnants of this can be related to non-ribosomal peptide synthesis which takes place largely in the cytoplasm of procaryotes.

A brief account of the events related to the above, should form an appropriate background to the present investigations and is outlined in the next section.

B. BACKGROUND

The very accurately working, but rather complex amino acid sequencing machinery, on the ribosome represents the final step in genetic information transfer. As such, it transmits from generation to generation information about the synthesis of specific proteins.

The protocols governing the ribosomal protein synthesis orchestrate a rather unusual chemical creation, composed as it is of various sequences of 20 different amino acids which have remained unchanged throughout the evolutionary process beginning with bacteria. Protein structure has the strange and useful property, depending on amino acid sequence, of giving rise to an inexhaustible variety of catalysts. An event, so complicated, so sophisticated and so precise could not have been born, so to say, in one fell swoop. The search for the "missing link" has led to the delineation of alternate pathways in peptide bond formation which take place largely in the cytoplasm of procaryotes and mediated by enzyme conglomerates.

Although the enzymatic formation of essential peptides such as glutathione and pantetheine was already known in the 'preribosomal era', the elucidation of the biosynthesis of more complex peptides followed the unravelling of the genetic code in the sixties[1−3]. A prediction of a poly- or multienzymatic pathway to peptides had been proposed as early as 1954[4] and a similar biosynthetic scheme of multienzymes as templates for polypeptides has now been verified for various types of peptides. Some of these studies have been reviewed. This background therefore attempts to present a more general view of this biosynthetic pathway.

If the mRNA directed ribosomal system is compared with a multienzymatic organization, one can observe the delicately controlled usage of the amino acids in the protein, manifested by the substrate recognition and correction events of tRNA aminoacylation by the aminoacyl-tRNA synthetases. This is in contrast to the liberal use by the pep-

tide synthetases of not only amino but hydroxy acids as substrate, the known number of carboxy-activated compounds at the moment being well over 300[5]. Both systems lead to linear peptide structures, but where as mRNAs coding for 2000 and more residues have been found, the longest peptide known to be formed by an enzyme system is still alamethicin (FIGURE.B.1), a 20-residue peptaibol[6].

Enzyme sub-structures have been found to be organized into multienzymes, and the protein template structure functions either as a set of interacting multienzymes or a single multienzymic structure. Which one of the the two organizational principles is used in each individual case cannot be predicted at present. So far no differently organized enzyme systems forming identical products have been characterized, in contrast to the case of fatty acid synthases. With a linear precursor, the amino terminal may be free, formylated, or acylated with a variety of carboxylic acids.

Peptide antibiotics that have been well studied, because of their important biological properties and which are synthesized on polyenzyme templates are presented in TABLE.B.1. Here the enzyme blocks are designated as Ez1, Ez2, ...Ezn. In all cases, the Ez1 initiates peptide synthesis and Ezn brings about termination.

Any of the polyenzymes that participate in gramicidin S (GS) or tyrocidine (Ty) synthesis may be separately charged with ATP and the corresponding amino acids (FIGURE.B.2 and FIGURE.B.3); however, these can be recovered by heating as single amino acids[14].

Filtration through sephadex G 200 separates from pre purified extracts of the strains of B. brevis fractions that catalyze the synthesis of gramicidin S. Two fractions, a light (GS1) and a heavy enzyme (GS2) of 100 x 10^3 and 280 x 10^3 molecular weight, that activate the amino acids inside the brackets in FIGURE.B.2 can easily be isolated.

The substrates are activated directly at their respective enzyme sites. There is no CoA-like activated form of the amino acids, as in polyketide formation. In the majority of cases studied so far, activation proceeds by cleavage of the α, β-phosphate bond of

Ac-Aib-Pro-Aib-Ala-Aib-Ala-Gln-Aib-Val-Aib-Gly-Leu-Aib-Pro-Val-Aib-Aib-Glu-Gln-Pheol

FIGURE.B.1

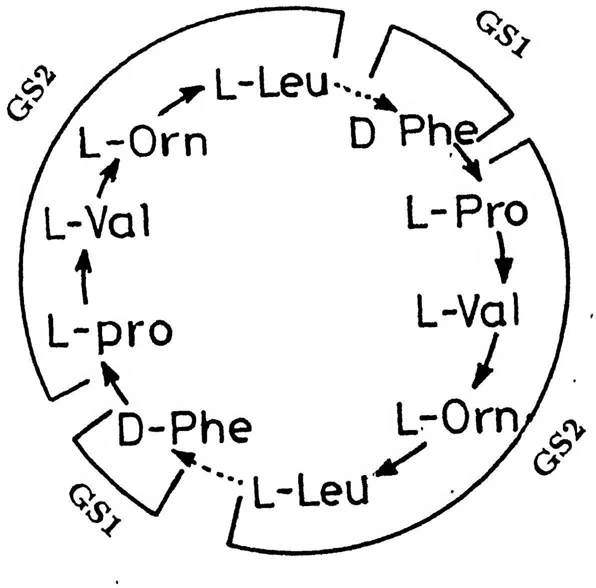


FIGURE.B.2

TABLE.B.1

Peptide antibiotics synthesized on polyenzyme templates

Peptide	Organism	Enzyme	Reference
Gramicidin	Bacillus brevis	LG1	
		LG2	7
		LGx	
Gramicidin S	Bacillus brevis	GS1	8
		GS2	
Tyrocidine	Bacillus brevis	TY1	
		TY2	9
		TY3	
Mycobacillin	Bacillus subtilis	MY1	
		MY2	10
		MY3	
Actinomycin	Streptomyces clavuligerus	AC1	
		AC2	11
		AC3	
Etamycin	Streptomyces griseus	ET1	12
		ETx	
Echinomycin	Streptomyces echinatus	EC1	
		EC2	12
		ECx	
Bacitracin	Bacillus licheniformis	BA1	
		BA2	13
		BA3	

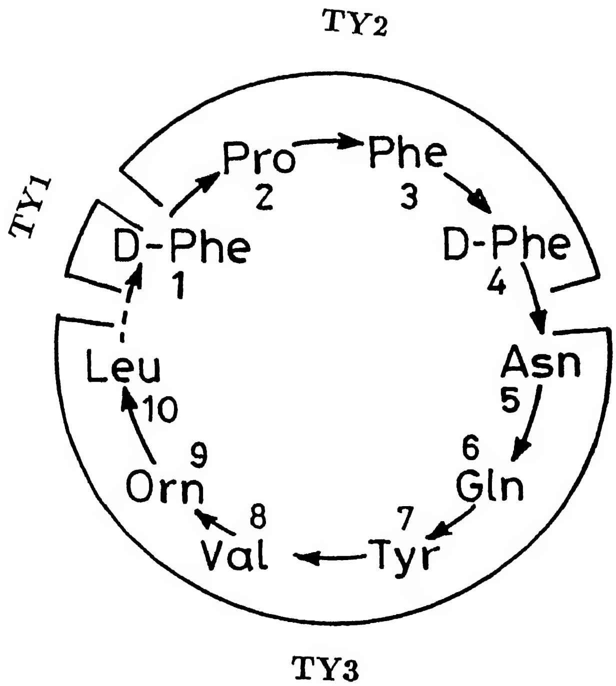


FIGURE.B.3

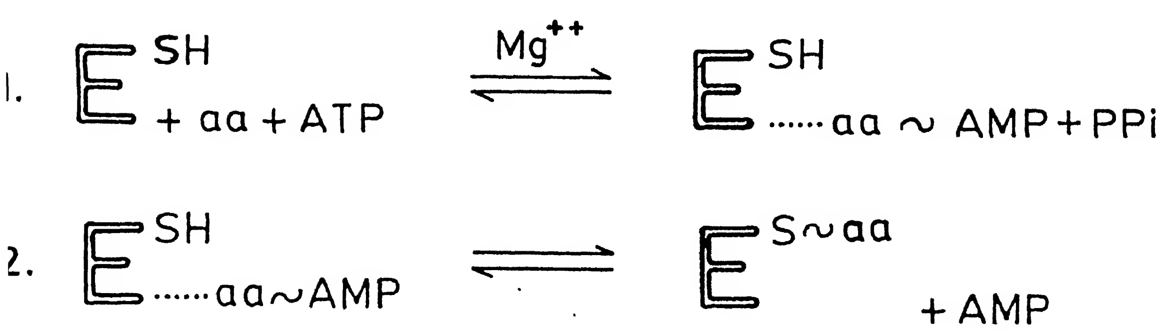


FIGURE. B.4

ATP, leading to the formation of an enzyme-stabilized, non-covalently bound amino or hydroxyl acyladenylate. Upon addition of pyrophosphate, the reaction is reversed and ATP is formed (FIGURE.B.4).

Sequence studies of the amino acid activation sites on GS2 tend to suggest that sequential conformational changes cause the movement of the Co-factor arm as shown in FIGURE.B.5 and FIGURE.B.6.

The biosynthesis of GS has been extensively studied. Significant results were obtained using ^{14}C labelled Phe. As could be seen from FIGURE.B.2, uptake of ^{14}C Phe would be continuous, provided all the other required amino acids are present. The ^{14}C activity profile is presented in TABLE.B.2.

Of great importance is the result of the last line in TABLE.B.2, namely, that when proline, the second amino acid, is omitted, but the following valine and ornithine are added, the enzyme bound [^{14}C] phenylalanine incorporation is back to the stage when phenylalanine alone was present. Thus, omission of one amino acid interrupts the progress of amino acid addition, similar to an amber mutation in ribosomal protein synthesis when an amino acid triplet in messenger RNA (mRNA) has undergone a mutation to a nonsense triplet[15]. Although the template is a different one here, interruption of the sequence similarly leads to premature termination.

Leucine addition rather rapidly releases gramicidin S coincident with disappearance of protein-bound peptides. Again, omission of amino acids causes polymerization to stop at the level reached before addition of the omitted amino acid. This confirms the interruption of vectorial amino acid polymerization by leaving out one in the sequence as seen in gramicidin S synthesis.

Thus, in case of gramicidin S formation, the synthetase 2 remains charged with di-, tri-, and tetrapeptide if the last amino acid is omitted.

A detailed study of tyrocidine [F*PFF*NQYVOrnL] biosynthesis have shown that the intermediates, F*P, F*PF, F*PFF* .... F*PFF*NQYVOrnL can be isolated and

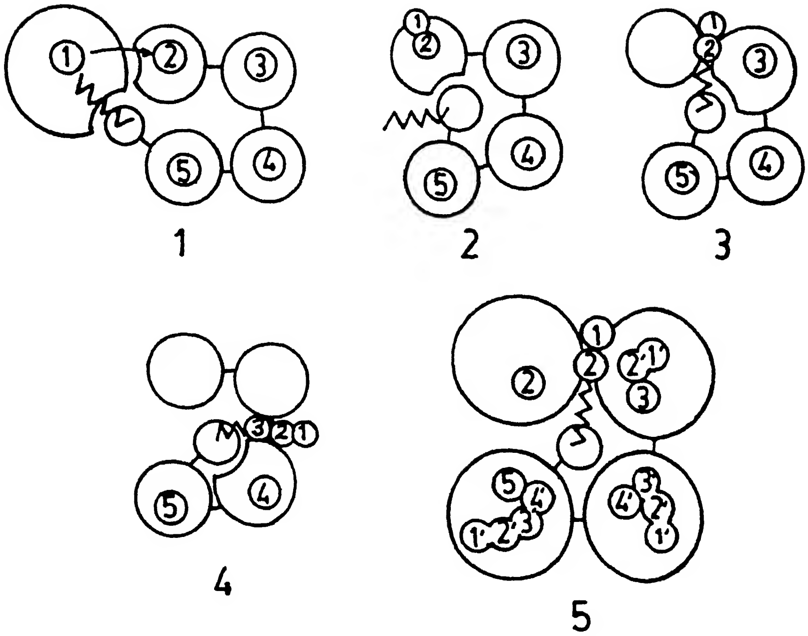


FIGURE.B.5

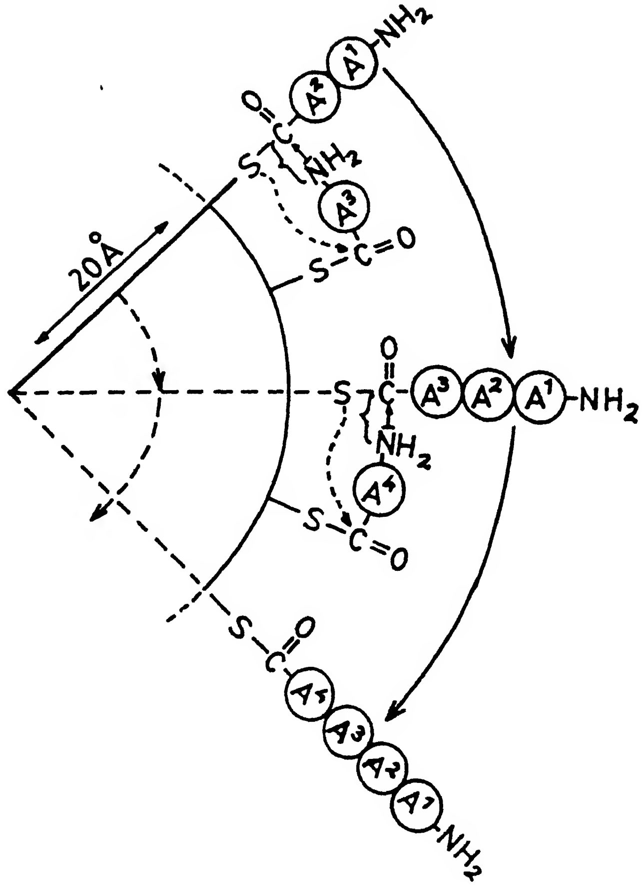


FIGURE.B.6

TABLE.B.2

Formation of protein-bound chains with increasing numbers of amino acids

Peptide chains formed	Protein-bound thioesters (count/min)	GS (count/min)
[^{14}C]Phe	1,549	265
[^{14}C]Phe-Pro	6,001	375
[^{14}C]Phe-Pro-Val	10,005	1,531
[^{14}C]Phe-Pro-Val-Orn	14,325	1,005
[^{14}C]Phe-Pro-Val-Orn-Leu	2,029	25,409
[^{14}C]Phe-....(Val, Orn)	1,610	376

characterized via artificial termination with CCl_3COOH.

The growing peptide chains are bound to the enzymes by thioester linkages. The requisite amino acids must be added in the proper order for polypeptide synthesis, if phenylalanine or proline are omitted, no peptide are formed. Omission of asparagine stops synthesis at the tetrapeptide stage, even when the succeeding amino acid present[16].

The above experiments clearly show that each enzyme unit is a precisely crafted template capable of generating a specific peptide module. Thus, the formation of non-ribosomal proteins can be considered as arising from a blockwise condensation mode.

It is tempting to speculate that the enzyme ensemble also evolved and that the initiation of the peptide synthesis and the need to have the assembly of all components to form products, also evolved from precursors that have ability to recognize specific amino acids in a predetermined arrangement.

Because of their therapeutic potential, the biosynthesis of several peptide antibiotics have been examined in recent years. All these have clearly shown that, regardless of their nature, the protocols of biosynthesis are same as is true of ribosomal protein synthesis. This aspect can be seen from SCHEMES.B.1, 2 and 3 that show the sequences involved in the biosynthesis of, respectively tyrocidine, bacitracin and the particularly illustrative linear Gramicidin-D. In these presentations, the enzyme blocks are differentiated by superscripts, thus permitting easy visualization of the peptide blocks and initiation and termination sites.

The lactone structures currently under biosynthetic investigation are representative members of different structural classes. Actinomycin (FIGURE.B.7), well known, highly toxic DNA-binding peptide formed by strains of streptomyces, are chromo- or acylpeptides. Actinomycin contains two simple lactone structures on a uniquely dimerizing chromophore. The biosynthesis of actinomycin is presented in SCHEME.B.4.

Non-ribosomal peptide synthesis also take place in eucaryotic organisms. The best illustration here would be that of cyclosporin (FIGURE.B.8) produced by the eucary-

SCHEME.B.1

Tyrocidine biosynthesis

(1)	Phe	$\rightarrow$	PheAMP	$\rightarrow$	E^1-S-DPhe	$\rightarrow$
(2)	Pro	$\rightarrow$	ProAMP	$\rightarrow$	E^2-S-Pro	$\rightarrow$
(3)	Phe	$\rightarrow$	PheAMP	$\rightarrow$	E^2-S-Phe	$\rightarrow$
(4)	Phe	$\rightarrow$	PheAMP	$\rightarrow$	E^2-S-DPhe	$\rightarrow$
(5)	Asn	$\rightarrow$	AsnAMP	$\rightarrow$	E^3-S-Asn	$\rightarrow$
(6)	Gln	$\rightarrow$	GlnAMP	$\rightarrow$	E^3-S-Gln	$\rightarrow$
(7)	Tyr	$\rightarrow$	TyrAMP	$\rightarrow$	E^3-S-Tyr	$\rightarrow$
(8)	Val	$\rightarrow$	ValAMP	$\rightarrow$	E^3-S-Val	$\rightarrow$
(9)	Orn	$\rightarrow$	OrnAMP	$\rightarrow$	E^3-S-Orn	$\rightarrow$
(10)	Leu	$\rightarrow$	LeuAMP	$\rightarrow$	E^3-S-Leu	$\rightarrow$

cyclization

$\Downarrow$

F*PFF*QYVOrnL

SCHEME.B.2

Bacitracin biosynthesis

(1)	Ile	$\rightarrow$	IleAMP	$\rightarrow$	E^1-S-Ile	$\rightarrow$
(2)	Cys	$\rightarrow$	CysAMP	$\rightarrow$	E^1-S-Cys	$\rightarrow$
(3)	Leu	$\rightarrow$	LeuAMP	$\rightarrow$	E^1-S-Leu	$\rightarrow$
(4)	Glu	$\rightarrow$	GluAMP	$\rightarrow$	E^1-S-DGlu	$\rightarrow$
(5)	Ile	$\rightarrow$	IleAMP	$\rightarrow$	E^1-S-Ile	$\rightarrow$
(6)	Lys	$\rightarrow$	LysAMP	$\rightarrow$	E^2-S-Lys	$\rightarrow$
(7)	Orn	$\rightarrow$	OrnAMP	$\rightarrow$	E^2-S-DOrn	$\rightarrow$
(8)	Ile	$\rightarrow$	IleAMP	$\rightarrow$	E^3-S-Ile	$\rightarrow$
(9)	Phe	$\rightarrow$	PheAMP	$\rightarrow$	E^3-S-DPhe	$\rightarrow$
(10)	His	$\rightarrow$	HisAMP	$\rightarrow$	E^3-S-His	$\rightarrow$
(11)	Asp	$\rightarrow$	AspAMP	$\rightarrow$	E^3-S-DAsp	$\rightarrow$
(12)	Asn	$\rightarrow$	AsnAMP	$\rightarrow$	E^3-S-Asn	$\rightarrow$

cyclization(Lys)

$\Downarrow$

ICLE*IKOrn*IF*HD*N

SCHEME.B.3

Gramicidin biosynthesis

formylation by tetrahydrofolate

(1)	Val	→	ValAMP	→	E^1-S-Val	→
(2)	Gly	→	GlyAMP	→	E^1-S-Gly	→
(3)	Ala	→	AlaAMP	→	E^2-S-Ala	→
(4)	Leu	→	LeuAMP	→	E^2-S-DLeu	→
(5)	Ala	→	AlaAMP	→	E^2-S-Ala	→
(6)	Val	→	ValAMP	→	E^2-S-DVal	→
(7)	Val	→	ValAMP	→	E^2-S-Val	→
(8)	Val	→	ValAMP	→	E^3-S-DVal	→
(9)	Trp	→	TrpAMP	→	E^*-S-Trp	→
(10)	Leu	→	LeuAMP	→	E^*-S-DLeu	→
(11)	Trp	→	TrpAMP	→	E^*-S-Trp	→
(12)	Leu	→	LeuAMP	→	E^*-S-DLeu	→
(13)	Trp	→	TrpAMP	→	E^*-S-Trp	→
(14)	Leu	→	LeuAMP	→	E^*-S-DLeu	→
(15)	Trp	→	TrpAMP	→	E^*-S-Trp	→

phosphatidyl-ethanolamine

⇓

ForVGAL*AV*VV*WL*WL*WL*WCH_2CH_2OH

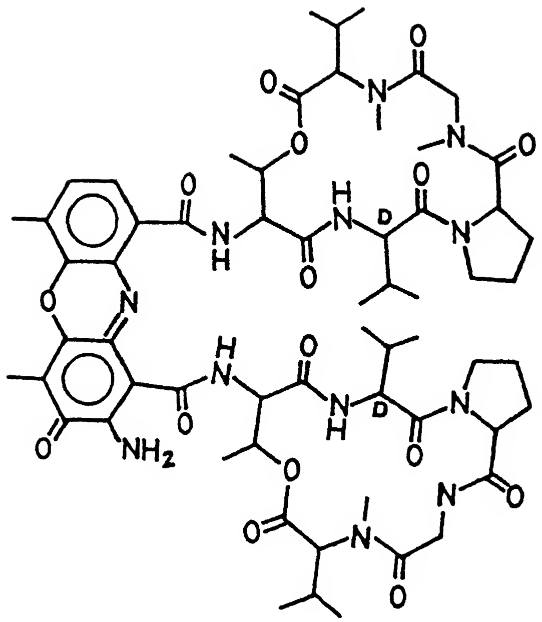


FIGURE.B.7

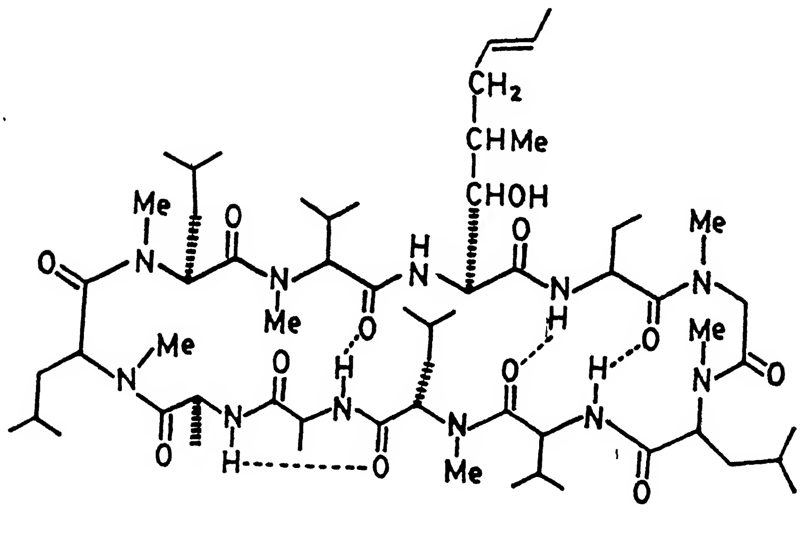


FIGURE.B.8

SCHEME.B.4

Biosynthesis of actinomycin

(A) Mha → E^1(MhaAMP) →

(1) Thr → ThrAMP → E^2-S-Thr →

(2) Val → ValAMP → E^2-S-DVal →

(3) Pro → ProAMP → E^3-S-Pro →

(4) Gly → GlyAMP → E^3-S-Gly → E^3-S-Sar →

(5) Val → ValAMP → E^3-S-Val → E^3-S-MeVal →

lactonization ↓

R-L-5

phenoxazinone formation ↓

$R^*(L\text{-}5)_2$

(Mha = 3-hydroxy-4-methyl anthranilic acid)

SCHEME.B.5

Biosynthesis of cyclosporin

(1)	DAla	→	DAlaAMP	→	E-S-DAla			→
(2)	Leu	→	LeuAMP	→	E-S-Leu	→	E-S-MeLeu	→
(3)	Leu	→	LeuAMP	→	E-S-Leu	→	E-S-MeLeu	→
(4)	Val	→	ValAMP	→	E-S-Val	→	E-S-MeVal	→
(5)	Bmt	→	BmtAMP	→	E-S-Bmt	→	E-S-MeBmt	→
(6)	Abu	→	AbuAMP	→	E-S-Abu			→
(7)	Gly	→	GlyAMP	→	E-S-Gly	→	E-S-Sar	→
(8)	Leu	→	LeuAMP	→	E-S-Leu	→	E-S-MeLeu	→
(9)	Val	→	ValAMP	→	E-S-Val			→
(10)	Leu	→	LeuAMP	→	E-S-Leu	→	E-S-MeLeu	→
(11)	Ala	→	AlaAMP	→	E-S-Ala			→

cyclization

otic Beauvenia nivea. As could be seen from the biosynthetic SCHEME.B.5, a single multifunctional enzyme orchestrate all the events.

Thus, in non-ribosomal peptide synthesis, large polyfunctional enzymes contain the activating enzymes arranged in sequence, that the thioester linked terminal of the growing chain transpeptidates from pantetheine to the preactivated peripheral amino acids and that the amino acid addition proceeds in a prescribed order until termination. The three dimensional set up of the polypeptide synthesis on a polyenzyme template remains to be resolved. So far, reliable attempts to detect organized structures by electron microscopy have been unsuccessful, nor has enough of the purified enzyme fractions been available to aim at crystallization.

It is very tempting to conclude that if chemical equivalents that can recognize and hold short stretches of amino acids and that in the event such units can be brought together, the chemical simulations of the biosynthesis of non-ribosomal peptides would be realized. The endeavours described in the following sections have a bearing on this theme.

C. PRESENT WORK

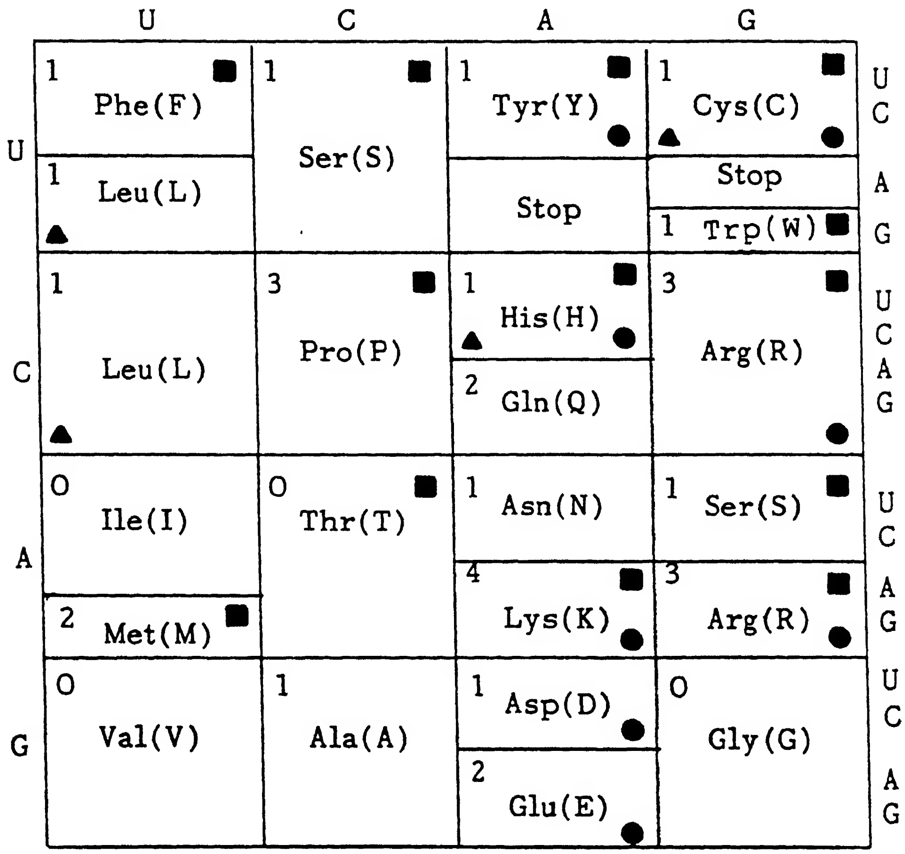
C.I. THE CORRELATION OF PROTEIN EVOLUTION WITH INTRINSIC PREFERENCES FOR PEPTIDE BOND FORMATION

Across the living domain, the functional system is uniformly comprised of 20 α-amino acids. Considering the fact that an infinite number of α-amino acid structures are possible and that well over 800 such α-amino acids do occur in nature, the selection in its code complement, the 20 amino acids must have met the rigours of evolution (CHART.C.I.1). Nearly six decades ago John Barnal had conjectured that, were these 20 α-amino acids be allowed to condense in water, the sequence of peptides and proteins thus produced will be non-random. This line of thought has always played a key role in the rationalization of possible pathways pertaining to the evolution of proteins, the formation of the information-function composite and culminating in the development of a protocell. The fact that the imprints of protein evolution are latent in present day proteins and enzymes is undisputed. Indeed the cornerstone for their construction are surely the non-randomly generated peptide segments solely arising from their structural profile. Yet, considering the fact that even today it is impossible to predict the peptide profile amongst four possibilities arising from a single peptide bond formation involving two amino acids, such endeavours, either involving more numbers of amino acids or pertaining to the primary sequences of larger segments, looks hopelessly complex. On the other hand our knowledge of protein structures and function would remain incomplete till a rational possibility can be arrived at involving the side chains of the coded amino acids and their proclivity for selective peptide bond formation.

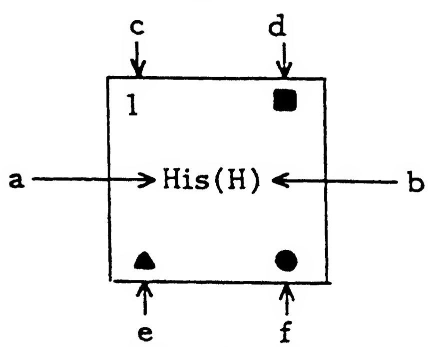
In the information system the sequence of code bases read in triplets can now be translated easily to a sequence of amino acids, but that would be a primary sequence of a protein or an enzyme. What would be most desirable would be the ability to predict the folding of such primary sequences to secondary and tertiary structures, i.e. the formation

CHART.C.I.1

GENETIC CODE



KEY



a. 3 letter abbreviation
b. 1 letter symbol
c. Spacer methylenes number
d. Oxidizable side chain
e. participant in templates
f. Capability to form trans-annular bridges.

of unique structural ensembles, precisely crafted to bring about a specific function. This aspect, rooted in the organic chemistry of the coded amino acid side chains, with focus on mutual interactions, is largely an unchartered domain. The understanding of protein folding has in recent times assumed great importance from a scientific angle. Such information is useful in protein design and from a practical perspective, in providing directions for genetically engineered proteins to fold according to precise protocols, since the lack of it would lead to irreversibly misfolded proteins of no use. A basic input here again, would be an understanding of the latent preference that surely exists in the coded complement for peptide formation. Thus the organic chemistry of coded amino acids side chains plays a pivotal role in the formation of protein structures.

Two broad objectives were sought in the present endeavours, namely, (a) to delineate the preference profile, if any, present in proteins and enzymes with respect to the choice of neighbours and (b) in the event it could be demonstrated that such preferences do exist, devise an experimental protocol to determine whether such preferences are intrinsic, i.e pertaining to the amino acids side chain mostly.

The preference profile present in proteins and enzymes with respect to the choice of neighbours necessarily involved the construction of a broad based data base of proteins. In order to make this analysis meaningful and versatile, proteins incorporated into the data base must have their structure established by high resolution x-ray diffraction analysis, so that not only are their three dimensional structures established but also the sequences of residues that are located in their secondary structure such as α-helices and β-sheets could be easily identified. Thus, the primary data base was constructed using sequences of proteins listed in TABLE.C.I.1. As could be seen from TABLE.C.I.1 each of the proteins in the set has been analyzed in terms of total no of residues present as well as number of residues which can be clearly seen either as an α-helix or a β-sheet.

A more detailed profile in terms of 20 coded amino acids that are present in the primary data base is shown in CHART.C.I.2. The listing here sequentially is, the total

TABLE.C.I.1

LISTING OF PROTEINS USED IN THE CONSTRUCTION OF THE DATA BASE

No.	Name	Residue No. Total	Residue No in α-Helix	Residue No in β-Sheet
1.	ACTINOXANTHIN	108	-	48
2.	CALCIUM BINDING PARVALBUMIN B	108	52	-
3.	CRAMBIN	46	21	8
4.	SUBTILISIN CARLSBERG E	274	93	49
5.	SUBTILISIN CARLSBERG I	70	18	20
6.	RIBOSOMAL PROTEIN	74	-	-
7.	HEMOGLOBIN (ERYTHROCRUORIN)	136	119	-
8.	IMMUNOGLOBULIN FAB (L)	216	-	-
9.	IMMUNOGLOBULIN FAB (H)	229	-	-
10.	FERREDOXIN	54	-	-
11.	FLAVODOXIN	148	-	-
12.	GAMMA-/II $ CRYSTALLIN	174	19	90
13.	OXIDISED IRON PROTEIN (HIPIP)	85	9	14
14.	HEMERYTHRIN (MET)	113	78	-
15.	INSULIN A	21	16	-
16.	INSULIN B	30	12	3
17.	MYOGLOBIN (DEOXY)	153	121	-
18.	MELITTIN	26	24	-
19.	PLASTOCYANIN	99	5	68
20.	AVIAN PANCREATIC POLYPEPTIDE	36	26	-
21.	BENCE- ⋆ JONES IMMUNOGLOBULIN	107	-	49
22.	RIBONUCLEASE	104	17	36
23.	SCORPION NEUROTOXIN (VARIANT 3)	65	10	-
24.	TONIN	235	-	-
25.	BETA-TRYPSIN	223	-	-

No.	Name	Residue No. Total	Residue No in α-Helix	Residue No in β-Sheet
26.	UBIQUITIN	76	16	33
27.	ACTINIDIN (SULFHYDRYL PROTEINASE)	220	65	32
28.	ALPHA-LYTIC PROTEASE	198	-	-
29.	ACID PROTEINASE, PENICILLOPEPSIN	323	48	170
30.	AZURIN (OXIDISED)	129	18	61
31.	CYTOCHROME B5 (OXIDISED)	93	51	25
32.	CYTOCHROME C=3	107	19	6
33.	CONCANAVALIN A	237	4	121
34.	CYTOCHROME P450 CAM	414	221	76
35.	CITRATE SYNTHASE	437	315	7
36.	CYTCHROME $ C PEROXYDASE	294	155	-
37.	ERABUTOXIN $ B	62	-	11
38.	HEMOGLOBIN V (CYANO, MET)	149	113	-
39.	LYSOZYME	164	107	18
40.	MYOHEMERYTHRIN	118	76	-
41.	OVOMUCOID THIRD DOMAIN	56	12	16
42.	PREALBULIN (HUMAN PLASMA)	127	9	56
43.	KALLIKREIN A UNIT A	80	-	-
44.	KALLIKREIN A UNIT B	152	-	-
45.	BENCE-⋆ JONES PROTEIN	114	8	51
46.	STAPHYLOCOCCAL NUCLEASE	149	34	32
47.	TRP REPRESSOR	107	78	-
48.	CYTOCHORME $C=551=(OXIDISED)	82	42	-
49.	CYTOCHROME $C=2=(REDUCED)	112	61	-
50.	DIHYDROFOLATE REDUCTASE	162	40	60
51.	GLUTATHIONE REDUCTASE	478	182	153
52.	RAT MAST CELL PROTEASE	224	-	-
53.	RUBREDOXIN	52	-	-

No.	Name	Residue No. Total	Residue No in α-Helix	Residue No in β-Sheet
54.	WHEAT GERM AGGLUTININ(ISOLECTIN 2)	170	23	-
55.	CARBOXYPEPTIDASE A	307	116	45
56.	HEMOGLOBIN A (DEOXY)	141	110	-
57.	HEMOGLOBIN B (DEOXY)	146	114	-
58.	LACTATE DEHYDROGENASE	229	145	63
59.	TROPONIN C	162	104	-
60.	TRYPSIN INHIBITOR	58	16	14
61.	RIBONULEASE A	124	32	58
62.	LYSOZYME	129	35	20

CHART.C.I.2

Coded amino acid occurrence in the data base (9416 residues) including those present in secondary structure elements

Amino Acid Residue	Total No. in Proteins	% Occurrence	Total No. in α-Helix	% Occurrence	Total No. in β-Sheet	% Occurrence
Alanine	815	8.65	330	10.96	100	6.60
Cysteine	252	2.67	52	1.72	38	2.51
Aspartic Acid	545	5.78	178	5.91	62	4.09
Glutamic Acid	495	5.25	219	7.27	56	3.70
Phenylalanine	354	3.75	138	4.58	74	4.89
Glycine	835	8.86	164	5.45	107	7.07
Histidine	239	2.53	94	3.19	35	2.31
Isoleucine	439	4.66	145	4.81	117	7.73
Lysine	579	6.14	216	7.17	77	5.08
Leucine	709	7.52	300	9.97	129	8.52
Methionine	172	1.82	77	2.55	31	2.04
Asparagine	421	4.47	118	3.92	46	3.04
Proline	427	4.53	74	2.45	34	2.24
Glutamine	349	3.70	124	4.12	58	3.83
Arginine	309	3.28	115	3.82	42	2.77
Serine	745	7.91	169	5.61	111	7.33
Threonine	586	6.22	157	5.21	116	7.66
Valine	656	6.96	204	6.77	176	11.63
Tryptophan	139	1.47	42	1.39	23	1.52
Tyrosine	350	3.71	93	3.09	81	5.35
Total	9416		3009		1513	

no present in 62 proteins (% occurrence), total in helices (%occurrence) and total in β-sheets (% occurrence). The sequences of residues totalling 9416 residues are presented as APPENDIX.C.I.1.

A comparison of the percentage occurrence in the general data set with that in α-helix and β-sheet is quite revealing. It can be seen from CHART.C.I.2 that the well recognized helix makers such as alanine, glutamic acid, leucine and lysine show substantial higher percentages, at the same time the residues which are generally detrimental to the stability of α-helix, such as glycine and proline show a substantial decrease. Thus, we see here not only a preference for neighbours but also an additional selectivity which is appropriate to the secondary structural element involved. Precisely, similar considerations apply in the β-sheet region. Coded amino acids which are highly prevalent in sheet structures such as isoleucine and valine exhibit a highly enhanced percentage of occurrence and those which are not favourable, such as, alanine, glutamic acid, lysine and leucine a significant lower value.

Based on computational protocol delineated in CHART.C.I.3, gross neighbour preferences involved in dipeptides for each of the 20 coded amino acids was generated. At the same time, for comparison purposes a parallel set was constructed assuming no preference in the choice of neighbour in the dipeptides using the following equation :

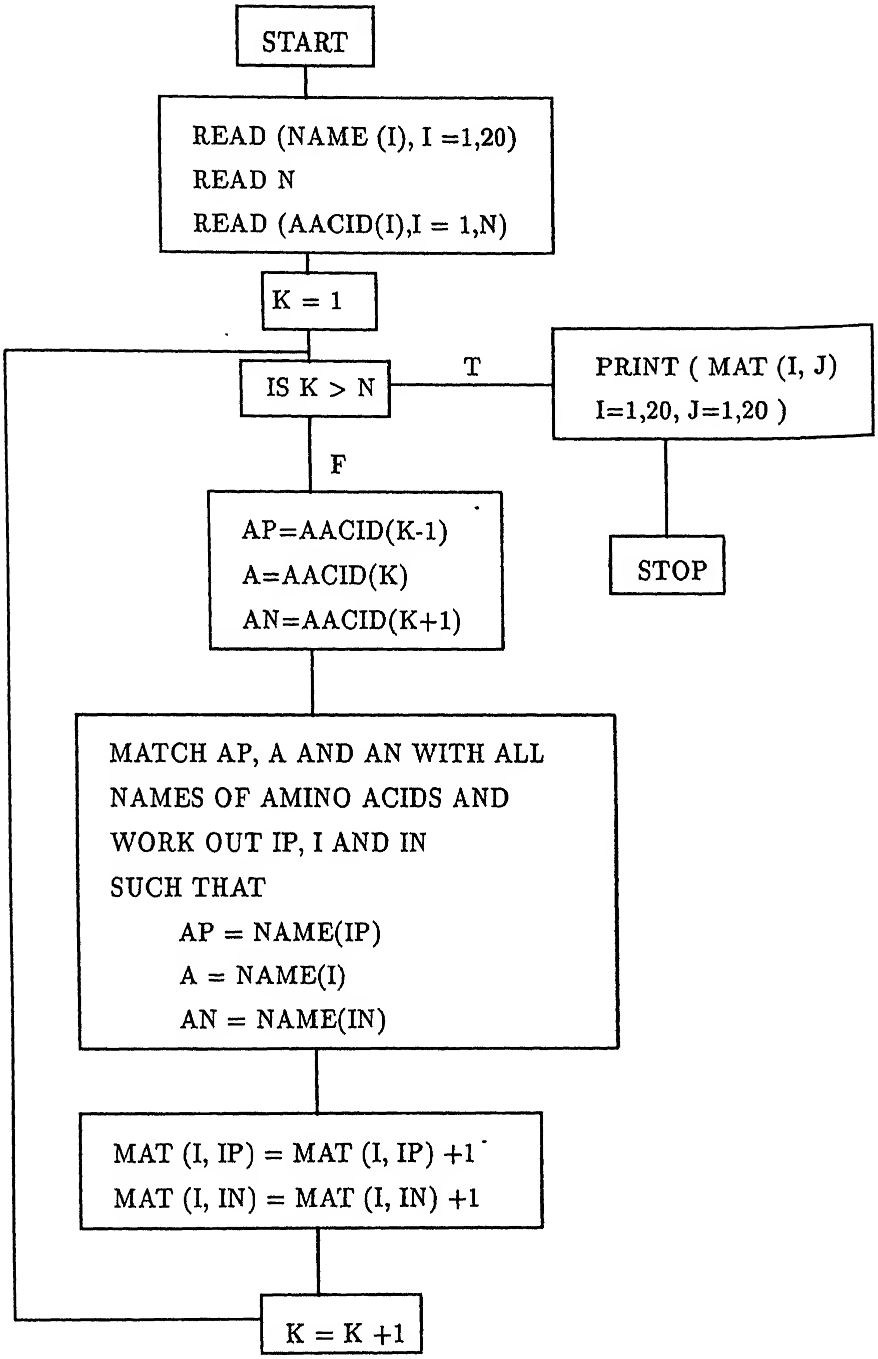
Non Preference Value = (2 x m x n) / total no of all amino acids residues present in whole set, where m and n are the numbers of respective amino acid residues present in the whole set.

A composite of these two operation is presented as a 20 by 20 grid in CHART.C.I.4 wherein relative neighbour preferences observed in the data set computed using flowchart described in CHART.C.I.3 and those expected on the basis of the above equation.

Even a cursory examination of CHART.C.I.4 would reveal that existence of preference for neighbours, a basic requirement for the presence of non random peptides, necessary when protein evolution on a blockwise approach is considered. The information present

CHART.C.I.3

Flowchart for Neighbour Residue Analysis



NAME : Array of dimension 20 which contains names of all amino acids (20)

N = Total number of amino acids present in proteins

AACID = Array of dimension N, stores distribution of N amino acids in proteins

MAT = Matrix of dimension 20 by 20. Value of MAT(I,J) represents number of occurrence when NAME(J) found to be NAME(I) of neighbour.

CHART.C.I.4

Relative neighbour preferences observed in a total of 9956 pairs of examples (top right) and those expected on the basis of non preference for neighbour (bottom left)

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	
	170	45	114	76	47	151	43	61	107	101	24	65	65	70	41	123	100	127	25	51	A
A	141	14	27	19	14	70	17	15	26	27	06	24	24	22	19	45	25	32	09	21	C
C	43	13	56	68	57	94	21	51	69	66	19	48	51	35	20	79	67	81	21	43	D
D	94	29	63	74	43	72	19	34	65	90	21	37	49	38	33	68	52	66	16	44	E
E	86	26	57	52	26	57	12	34	47	48	11	32	30	23	31	61	63	40	03	27	F
F	61	19	19	37	27	160	44	80	88	104	26	45	69	63	64	173	97	108	25	66	G
G	144	45	97	88	63	148	6	22	36	50	07	15	37	20	14	37	27	32	05	15	H
H	41	13	28	25	18	42	12	58	50	70	14	34	45	31	34	62	65	64	19	28	I
I	76	23	51	46	33	78	22	41	68	108	32	46	48	27	25	90	75	85	10	48	K
K	100	31	76	61	43	103	29	54	71	108	31	68	65	53	67	110	87	105	17	36	L
L	123	38	82	74	53	127	36	66	87	107	16	16	14	15	12	23	13	25	05	09	M
M	30	09	20	18	13	30	09	16	21	26	07	56	44	23	28	70	62	65	15	39	N
N	73	22	49	44	32	75	21	39	52	63	15	38	28	31	24	69	57	55	11	35	P
P	74	23	49	45	32	76	22	40	52	64	16	38	39	42	33	55	33	35	12	33	Q
Q	60	19	40	37	26	62	18	33	43	52	13	31	32	26	24	37	28	51	09	19	R
R	53	17	36	32	23	55	16	29	38	46	11	28	28	23	20	112	92	93	24	61	S
S	129	40	86	78	56	132	38	69	91	112	27	67	67	55	49	118	60	92	16	53	T
T	101	31	68	62	44	104	30	55	72	88	21	52	53	43	38	93	73	92	25	34	V
V	113	35	76	67	49	116	33	61	81	99	24	59	59	49	43	104	82	91	02	09	W
W	24	07	16	15	10	25	07	13	17	21	05	12	13	10	09	22	17	19	04	28	Y
Y	60	19	40	37	26	62	18	33	43	53	13	31	32	26	23	55	44	49	10	26	
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	

COMPUTED VALUE

NON PREFERENCE VALUE

in CHART.C.I.4 is exhibited for each of the 20 coded amino acids wherein % deviation in the preference for neighbour is clearly shown (CHART.C.I.5).

Obviously, each of the coded amino acid offers two sites for peptide bond formation, either involving the amino group or the carboxyl function. In the event peptidation occurs involving the α-amino group, the incoming residue would be placed at the amino end of the dipeptide or to the left of the central residue. On the other hand if the peptidation were to involve the carboxyl end, the incoming unit would be placed at the carboxyl end of the dipeptide or to the right of the central residue. To assess the importance of these two pathways with reference to peptide bond formation in dipeptides, another computer program, as shown in flowchart (CHART.C.I.6) was employed.

In CHART.C.I.7, is presented a master analysis for neighbour which also indicate the "Left-Right" preference. These are derived from major data base consisting of 62 proteins and 9416 residues *vide supra*. CHART.C.I.7 presents many interesting features. A superficial analysis of this would immediately show that the left-right preferences are unequal amongst all in all the 20 coded amino acids. CHART.C.I.7 also turned out to be an important indicator with reference to the construction of dipeptide and higher sequences. CHART.C.I.4 in conjunction with CHART.C.I.7 support the existence of neighbour preference as well as preference pertaining to the placement of the neighbour residue. Thus, on this basis alone the notion of the existence of peptides of specific sequences in the early stages of protein evolution becomes very logical. An outcome of such an analysis would be the expectation that the neighbour preferences as well as the left-right preferences would reflect a general profile of a secondary peptide structural element such as the β-sheet or the α-helix, since developments in this domain have clearly shown that the coded amino acid residues which are generally associated with promotion of such secondary structural elements are different [17–19]. As could be seen from TABLE.C.I.1, of the 9416 residues present in the 62 proteins, it can be found from x-ray crystallographic data, that of these, 3009 residues in 47 proteins can be placed as

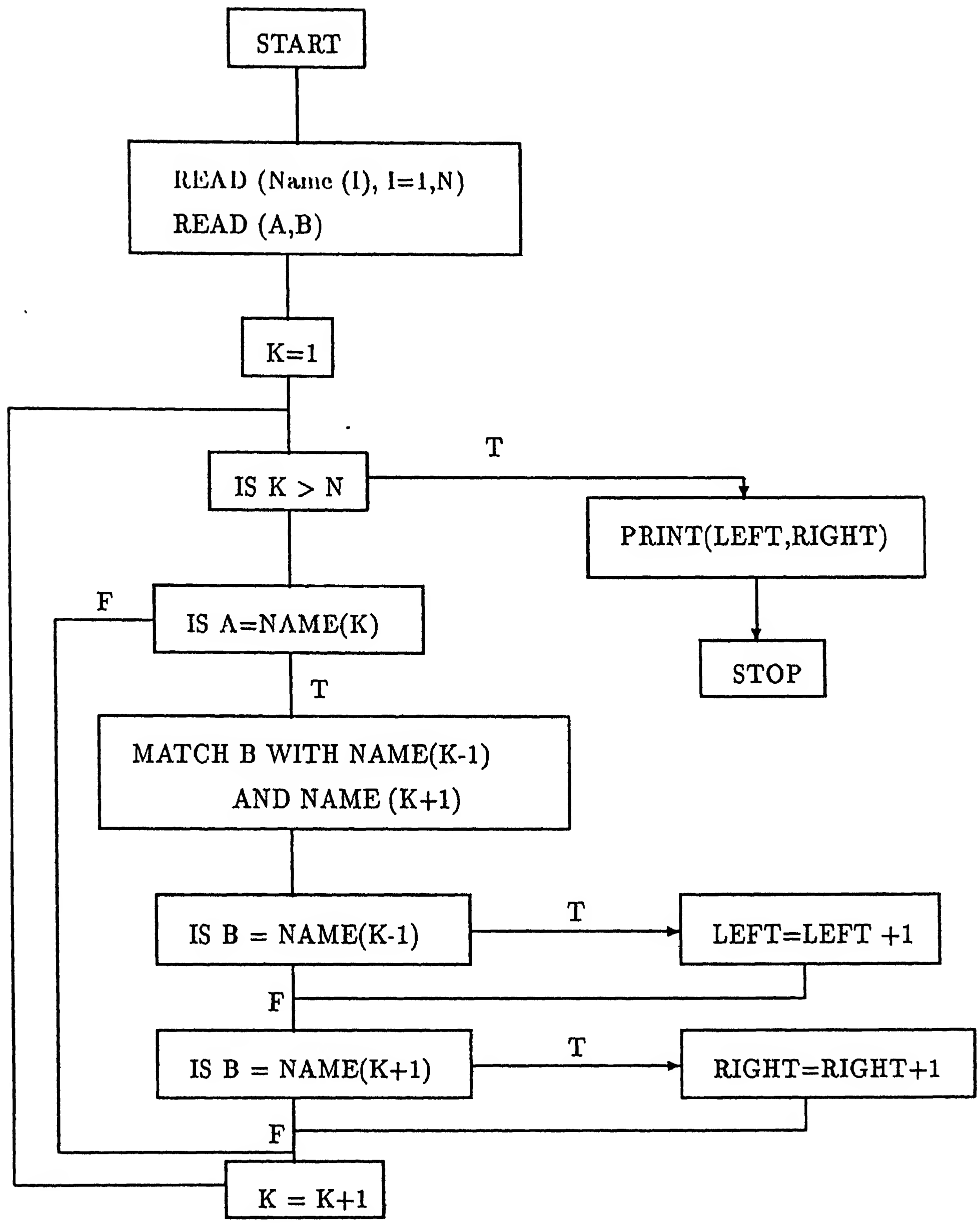
CHART.C.I.5

Percentage deviation in neighbour preferences observed in a total of 9956 pairs from those expected on the basis of non preference for neighbours

A	+21	+05	+21	-12	-23	+05	+05	-20	+07	-18	-20	-11	-12	+17	-23	-05	-01	+12	+04	-15
C	+05	+08	-07	-27	-26	+56	+31	-35	-16	-29	-33	+09	+04	+16	+12	+13	-19	-09	+29	+11
D	+21	-07	-11	+19	+39	-03	-25	00	+03	-20	-05	-02	+04	-13	-44	-08	-01	+07	+31	+08
E	-12	-27	+19	+42	+14	-18	-24	-26	+07	+22	+17	-16	+09	+03	+03	-13	-16	-01	+07	+19
F	-23	-26	+39	+14	-04	-10	-33	+03	+09	-09	-15	00	-06	-12	+35	+09	+43	-18	-70	+04
G	+05	+56	-03	-18	-10	+08	+05	+03	-15	-18	-13	-40	-09	+02	+16	+31	-07	-07	00	+06
H	+05	+31	-25	-24	-33	+05	-50	00	+24	+39	-22	-29	+68	+11	-13	-03	-10	-03	-29	-17
I	-20	-35	00	-26	+03	+03	00	+41	-07	+06	-13	-13	+13	-06	+17	-10	+18	+05	+46	-15
K	+07	-16	+03	+07	+09	-15	+24	-07	-04	+24	+52	-12	-08	-37	-34	-01	+04	+05	-41	+12
L	-18	-29	-20	+22	-09	-18	+39	+06	+24	+01	+19	+08	+02	+02	+46	-02	-01	+06	-19	-32
M	-20	-33	-05	+17	-15	-13	-22	-13	+52	+19	+129	+07	-13	+15	+09	-15	-38	+04	00	-31
N	-11	+09	-02	-16	00	-40	-29	-13	-12	+08	+07	+47	+16	-26	00	+04	+19	+10	+25	+26
P	-12	+04	+04	+09	-06	-09	+68	+13	-08	+02	-13	+16	-28	-03	-14	+03	+08	-07	-15	+09
Q	+17	+16	-13	+03	-12	+02	+11	-06	-37	+02	+15	-26	-03	+62	+43	00	-23	-29	+20	+27
R	-23	+12	-44	+03	+35	+16	-13	+17	-34	+46	+09	00	-14	+43	+20	-24	-26	+19	00	-17
S	-05	+13	-08	-13	+09	+31	-03	-10	-01	-02	-15	+04	+03	00	-24	-05	-01	-11	+09	+11
T	-01	-19	-01	-16	+43	-07	-10	+18	+04	-01	-38	+19	+08	-23	-26	-01	-18	+12	-06	+20
V	+12	-09	+07	-01	-18	-07	-03	+05	+05	+06	+04	+10	-07	-29	+19	-11	+12	+01	+36	-31
W	+04	+29	+31	+07	-70	00	-29	+46	-41	-19	00	+25	-15	+20	00	+09	-06	+36	-50	-10
Y	-15	+11	+08	+19	+04	+06	-17	-15	+12	-32	-31	+26	+09	+27	-17	+11	+20	-31	-10	+08
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y

CHART.C.I.6

Flowchart for "Left-Right" Preference Analysis



NAME : Array of dimension n containing the protein sequence

N : Total number of amino acids in the protein sequence

A,B : Amino acids B is the residue to be tested for presence on the left or right of A

LEFT = Number of times B comes to the left of A

RIGHT = Number of times B comes to the right of A

CHART.C.I.7

MASTER ANALYSIS FOR NEIGHBOUR (LEFT-RIGHT) PREFERENCE

(62 Proteins, 9416 Residues)

NEIGHBOUR	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	85/85	21/24	56/58	32/44	24/23	78/73	27/16	34/27	52/55	55/46	15/09	30/35	32/33	36/34	21/20	61/62	51/49	64/63	12/13	23/28
C	24/21	07/07	10/17	09/10	04/10	41/29	06/11	07/08	14/12	12/15	04/02	17/07	14/10	13/09	09/10	23/22	09/16	13/19	06/03	07/14
D	58/56	17/10	28/28	32/36	32/25	50/44	06/15	27/24	31/38	37/29	04/15	22/26	23/28	15/20	10/10	47/32	35/32	31/50	08/13	30/13
E	44/32	10/09	36/32	37/37	23/20	37/35	09/10	12/22	29/36	51/39	14/07	21/16	17/32	22/16	15/18	33/33	22/30	32/34	12/04	18/26
F	23/24	10/04	25/32	20/23	13/13	21/36	03/09	16/18	27/20	25/23	06/05	16/16	23/07	11/12	18/13	31/30	32/31	16/24	02/01	15/12
G	73/78	29/41	44/50	35/37	36/21	80/80	24/20	48/32	48/40	53/51	15/11	22/23	25/44	29/34	35/29	80/93	54/43	55/53	09/16	34/32
H	16/27	11/06	15/06	10/09	09/03	20/24	03/03	09/13	22/14	28/22	02/05	05/10	22/15	07/13	06/08	16/21	14/13	16/16	02/03	05/08
I	27/34	08/07	24/27	22/12	18/16	32/48	13/09	29/29	26/24	25/45	07/07	18/16	32/13	15/16	13/21	29/33	35/30	38/26	07/12	19/09
K	55/52	12/14	38/31	36/29	20/27	40/48	14/22	24/26	34/34	54/54	15/17	27/19	19/29	11/16	14/11	53/37	32/43	44/41	03/07	29/19
L	46/55	15/12	29/37	39/51	23/25	51/53	22/28	45/25	54/54	54/54	13/18	33/35	42/23	30/23	37/30	59/51	46/41	47/58	06/11	14/22
M	09/15	02/04	15/04	07/14	05/06	11/15	05/02	07/07	17/15	18/13	08/08	08/08	07/07	07/08	07/05	10/13	07/06	13/12	03/02	05/04
N	35/30	07/17	26/22	16/21	16/16	23/22	10/05	16/18	19/27	35/33	08/08	28/28	20/24	12/11	11/17	32/38	32/30	38/27	08/07	19/20
P	33/32	10/14	28/23	32/17	07/23	44/25	15/22	13/32	29/19	23/42	07/07	24/20	14/14	11/20	08/16	44/25	25/32	33/22	09/02	16/19
Q	34/36	09/13	20/15	16/22	12/11	34/29	13/07	16/15	16/11	23/30	08/07	11/12	20/11	21/21	16/17	22/33	16/17	17/18	07/05	14/19
R	20/21	10/09	10/10	18/15	13/18	29/35	08/06	21/13	11/14	30/37	05/07	17/11	16/08	17/16	12/12	18/19	13/15	28/23	04/05	07/12
S	62/61	22/23	32/47	35/33	30/31	93/80	21/16	33/29	37/53	51/59	13/10	38/32	25/44	33/22	19/18	56/56	47/45	44/49	15/09	35/26
T	49/51	16/09	32/35	30/22	31/32	43/54	13/14	30/35	43/32	41/46	06/07	30/32	32/25	17/16	15/13	45/47	30/30	45/47	10/06	26/27
V	63/64	19/13	50/31	34/32	24/16	53/55	16/16	26/38	41/44	58/47	12/13	27/38	22/33	18/17	23/28	49/44	47/45	46/46	10/15	16/18
W	13/12	03/06	13/08	04/12	01/02	16/09	03/02	12/07	07/03	11/06	02/03	07/08	02/09	05/07	05/04	09/15	06/10	15/10	01/01	04/05
Y	28/23	14/07	13/30	26/18	12/15	32/34	08/05	09/19	19/29	22/14	04/05	20/19	19/16	19/14	12/07	26/35	27/26	18/16	05/04	14/14

CENTRAL RESIDUE

Read out format: Central residue X can have neighbour profile YX(left preference) or XY(right preference) example for P, FP:PF :: 23:07

α-helices and 1513 residues in 33 proteins in the β-sheet region. Neighbourhood preferences and left-right preferences pertaining to these two important secondary structural elements were performed in tandem and the results are presented in CHART.C.I.8 and CHART.C.I.9. A comparison of CHART.C.I.7 with CHART.C.I.8 and CHART.C.I.9 would tend to show that the left-right preferences are more pronounced, both in the case of α-helices and β-sheets compared to the general data set.

Significant left-right neighbour preferences presented in CHART.C.I.7 have been sorted out and presented in CHART.C.I.10. A general examination of CHART.C.I.10 would tend to show that the left-right preferences are highly pronounced when the central residue involved here is proline. This aspect has been highlighted in CHART.C.I.11 wherein proline preference profile is fitted into the genetic code format. In this representation not only the left-right preferences with respect to each of the 20 coded amino acids for dipeptide formation with proline is shown, but also the preference for proline as a neighbour in terms of percentage deviation (CHART.C.I.5). Thus CHART.C.I.11 not only shows that the preferences for peptide bond formation with proline in a selective manner is pervasive but also that the neighbour selectivity is accentuated. The most noteworthy aspect is the highly unfavourable profile with respect to the formation of Pro-Pro peptide bond which shows a very negative (-28 %) deviation from the non preference value. Further, CHART.C.I.11 also shows that none of the dipeptides involving proline show a random value of 0 % and the range of preference extends from -28 % for Pro to +68 % for histidine.

Whilst CHART.C.I.11 do reflect preferences, monitored in terms of a single coded amino acid, namely, proline, a general analysis pertaining to this appears too complex. This could be exemplified either in terms of homologous pairs, Q (11-20) and N (24-20) or in terms of isosteric systems V (33-22) and T (25-32). In view of this, a modest approach was chosen restricting to the understanding of the preference profile amongst the coded amino acids lacking functional groups in the side chains. The optimism here

CHART.C.I.8

MASTER ANALYSIS FOR NEIGHBOUR (LEFT-RIGHT) PREFERENCE IN

α-HELIX

(47 Proteins, 3009 Residues)

NEIGHBOUR	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	39/39	04/04	21/19	19/27	12/12	19/12	10/11	16/12	29/21	32/29	06/04	07/12	03/05	17/14	10/10	11/23	17/16	24/29	06/07	11/08
C	04/04	01/01	01/05	02/02	01/03	09/05	02/02	02/02	02/02	03/00	02/01	04/02	00/01	02/01	01/03	03/01	03/03	06/06	01/00	00/05
D	19/21	05/01	09/09	13/17	11/05	07/05	02/03	10/05	12/11	13/12	02/05	04/04	08/04	06/07	03/02	08/08	05/08	15/15	03/02	11/04
E	27/19	02/02	17/13	20/20	12/08	06/11	06/02	04/09	13/16	27/18	07/04	05/07	02/11	15/09	05/11	07/12	07/13	10/12	07/01	05/06
F	12/12	03/01	05/11	08/12	08/08	02/08	01/03	06/09	17/10	13/09	03/03	04/11	03/02	05/04	08/06	06/06	12/09	05/03	01/01	02/03
G	12/19	05/09	05/07	11/06	08/02	04/04	03/10	11/04	07/05	12/18	04/03	02/04	03/01	03/05	03/05	06/12	06/05	11/10	00/04	03/07
H	11/10	02/02	03/02	02/06	03/01	10/03	01/01	02/05	12/07	10/08	01/03	01/04	03/03	01/06	03/03	06/08	04/05	05/02	00/02	02/02
I	12/16	02/02	05/10	09/04	09/06	04/11	05/02	11/11	10/10	12/18	03/04	04/02	08/04	04/04	06/11	07/08	06/06	12/05	02/02	04/03
K	21/29	02/02	11/12	16/13	10/17	05/07	07/12	10/10	16/16	21/23	06/11	09/04	03/05	03/05	07/04	16/09	09/09	16/16	01/02	07/04
L	29/32	00/03	12/13	18/27	09/13	18/12	08/10	18/12	23/21	22/22	08/06	16/13	07/03	13/12	16/11	21/12	14/16	22/24	02/09	07/12
M	04/06	01/02	05/02	04/07	03/03	03/04	03/01	04/03	11/06	06/08	06/06	04/02	01/03	01/04	03/04	03/06	01/01	03/04	01/01	02/01
N	12/07	02/04	04/04	07/05	11/04	04/02	04/01	02/04	04/09	13/16	02/04	02/02	04/06	07/03	04/05	05/05	03/07	14/11	02/02	02/03
P	05/03	01/00	04/08	11/02	02/03	01/03	03/03	04/08	05/03	03/07	03/01	06/04	01/01	01/01	03/01	05/03	04/05	08/03	01/00	01/02
Q	14/17	01/02	07/06	09/15	04/05	05/03	06/01	04/04	05/03	12/13	04/01	03/07	01/01	09/09	08/09	01/08	03/04	06/06	03/01	04/05
R	10/10	03/01	02/03	11/05	06/08	05/03	03/03	11/06	04/07	11/16	04/03	05/04	01/03	09/08	07/07	04/06	05/03	04/09	02/01	03/03
S	23/11	01/03	08/08	12/07	06/06	12/06	08/06	08/07	09/16	12/21	06/03	05/05	03/05	08/01	06/04	05/05	08/09	05/11	02/03	06/04
T	16/17	03/03	08/05	13/07	09/12	05/06	05/04	06/06	09/09	16/14	01/01	07/03	05/04	04/03	03/05	09/08	05/05	13/12	01/01	07/04
V	29/24	06/06	15/15	12/10	03/05	10/11	02/05	05/12	16/16	24/22	04/03	11/14	03/08	06/06	09/04	11/05	12/13	12/12	03/02	03/04
W	07/06	00/01	02/03	01/07	01/01	04/00	02/00	02/02	02/01	09/02	01/01	02/02	00/01	01/03	01/02	03/02	01/01	02/03	00/00	01/02
Y	08/11	05/00	04/11	06/05	03/02	07/03	02/02	03/04	04/07	12/07	01/02	03/02	02/01	05/04	03/03	04/06	04/07	04/03	02/01	04/04
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y

CENTRAL RESIDUE

Read out format: Central residue X can have neighbour profile YX(left preference) or XY(right preference) example for P, FP:PF :: 03:02

CHART.C.I.9

MASTER ANALYSIS FOR NEIGHBOUR (LEFT-RIGHT) PREFERENCE IN β-SHEET

(33 Proteins, 1513 residues)

NEIGHBOUR	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	07/07	02/04	04/06	01/02	08/03	05/05	02/01	04/10	02/02	08/05	00/01	02/01	04/01	04/03	00/02	08/07	07/09	12/13	00/02	02/05
C	04/02	01/01	00/04	02/02	01/01	01/01	00/01	01/01	00/00	02/02	00/00	02/03	02/01	04/02	02/01	01/02	01/05	02/01	01/00	03/03
D	06/04	04/00	00/00	02/03	03/04	04/05	00/00	03/05	00/01	06/04	00/00	00/00	00/01	01/02	01/00	05/01	05/03	05/06	01/00	05/04
E	02/01	02/02	03/02	01/01	02/03	02/05	00/00	04/03	01/02	07/05	02/00	00/00	00/00	00/00	02/01	03/04	02/01	12/05	03/00	03/08
F	03/08	01/01	04/03	03/02	02/02	04/03	01/02	02/04	01/02	07/02	02/01	03/02	00/00	00/01	02/01	08/09	08/09	05/12	00/00	06/04
G	05/05	01/01	05/04	05/02	03/04	07/07	04/00	10/05	04/03	07/05	01/00	02/00	01/02	05/04	06/02	04/08	07/05	06/13	00/00	02/06
H	01/02	01/00	00/00	00/00	02/01	00/04	00/00	06/02	02/01	03/04	01/00	02/00	01/04	01/01	00/00	00/01	02/01	05/06	01/00	00/00
I	10/04	01/01	05/03	03/04	04/02	05/10	02/06	10/10	03/07	08/11	01/02	03/02	01/03	02/06	04/05	06/08	08/10	12/09	02/03	07/03
K	02/02	00/00	01/00	02/01	02/01	03/04	01/02	07/03	04/04	10/07	02/02	02/02	01/01	02/01	01/01	03/04	10/05	11/09	01/01	04/01
L	05/08	02/02	04/06	05/07	02/07	05/07	04/03	11/08	07/10	16/16	03/01	03/04	03/04	02/05	03/06	08/08	04/07	10/12	03/01	04/02
M	01/00	00/00	00/00	00/02	01/02	00/01	00/01	02/01	02/02	01/03	00/00	00/01	01/01	03/01	02/01	03/01	02/02	07/03	01/01	01/01
N	01/02	03/02	00/00	00/00	02/03	00/02	00/02	02/03	02/02	04/03	01/00	01/01	00/01	00/01	01/01	02/03	04/05	08/02	02/01	00/02
P	01/04	01/02	01/00	00/00	00/00	02/01	04/01	03/01	01/01	04/03	01/01	01/00	00/00	02/01	00/02	03/01	01/02	05/04	00/01	02/00
Q	03/04	02/04	02/01	00/00	01/00	04/05	01/01	06/02	01/02	05/02	01/03	01/00	01/02	05/05	02/01	05/02	04/02	02/02	00/01	04/03
R	02/00	01/02	00/01	01/02	01/02	02/06	00/00	05/04	01/01	06/03	01/02	01/01	02/00	01/02	01/01	01/01	02/02	08/02	00/00	01/00
S	07/08	02/01	01/05	04/03	09/08	08/04	01/00	08/06	04/03	08/08	01/03	03/02	01/03	02/05	01/01	03/03	06/05	12/14	01/03	06/06
T	09/07	05/01	03/05	01/02	09/08	05/07	01/02	10/08	05/10	07/04	02/02	05/04	02/01	02/04	02/02	05/06	07/07	09/09	01/01	06/08
V	13/12	01/02	06/05	05/12	12/05	13/06	06/05	09/12	09/11	12/10	03/07	02/08	04/05	02/02	02/08	14/12	09/09	19/19	02/04	04/03
W	02/00	00/01	00/01	00/03	00/00	00/00	00/01	03/02	01/01	01/03	01/01	01/02	01/00	01/00	00/00	03/01	01/01	04/02	00/00	02/01
Y	05/02	03/03	04/05	08/03	04/06	06/02	00/00	03/07	04/04	02/04	01/01	02/00	00/02	03/04	00/01	06/06	08/06	03/04	01/02	06/06

CENTRAL RESIDUE

Read out format: Central residue X can have neighbour profile YX(left preference) or XY(right preference) example for P, GP:PG :: 01:02

CHART.C.I.10

SIGNIFICANT NEIGHBOUR (LEFT-RIGHT) PREFERENCE

(62 Proteins, 9416 Residues)

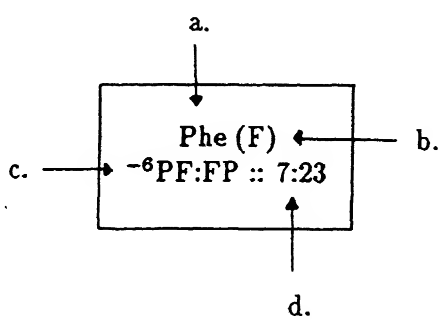
NEIGHBOUR	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A				32/44			27/16	34/27		55/46	15/09									
C			10/17			41/29						17/07					09/16			07/14
D		17/10					06/15				04/15					47/32		31/50		30/13
E	44/32							12/22		51/39	14/07		17/32						12/04	
F						21/36			27/20				23/07					16/24		
G		29/41			36/21			48/32					25/44				54/43		09/16	
H	16/27		15/06						22/14				22/15							
I	27/34			22/12		32/48				25/45			32/13		13/21			38/26		19/09
K					20/27		14/22					27/19	19/29			53/37	32/43			29/19
L	46/55			39/51				45/25					42/23							
M	09/15		15/04	07/14																
N		07/17							19/27									38/27		
P				32/17	07/23	44/25	15/22	13/32	29/19	23/42				11/20	08/16	44/25	25/32			
Q													20/11			22/33				
R								21/13					16/08							
S			32/47						37/53				25/44	33/22					15/09	
T		16/09				43/54			43/32				32/25						10/06	
V			50/31		24/16			26/38				27/38								
W				04/12		16/09										09/15	06/10			
Y		14/07	13/30					09/19	19/29											

CENTRAL RESIDUE

Read out format: Central residue X can have neighbour profile YX(left preference) or XY(right preference) example for P, FP:PF :: 23:07

CHART.C.I.11
PROFILE OF PROLINE (P) DIPEPTIDES

	U	C	A	G	
U	Phe (F) $^{-6}$PF:FP :: 7:23 Leu (L)	Ser (S) $^{+3}$PS:SP :: 44:25	Tyr (Y) $^{+9}$PY:YP :: 16:19 Term	Cys (C) $^{+4}$PC:CP :: 10:14 Term Trp (W) $^{-15}$PW:WP :: 9:2	U C A G
C	Leu (L) $^{+2}$PL:LP :: 23:42	Pro (P) $^{-28}$PP:PP :: 14:14	His (H) $^{+68}$PH:HP :: 15:22 Gln (Q) $^{-3}$PQ:QP :: 11:20	Arg (R) $^{-14}$PR:RP :: 8:16	U C A G
A	Ile (I) $^{+13}$PI:IP :: 13.32 Met (M) $^{-3}$PM:MP :: 7:7	Thr (T) $^{+8}$PT:TP :: 25:32	Asn (N) $^{+16}$PN:NP :: 24:20 Lys (K) $^{-8}$PK:KP :: 29:19	Ser (S) Arg (R)	U C A G
G	Val (L) $^{-7}$PV:VP :: 33:22	Ala (A) $^{-12}$PA:AP :: 33:32	Asp (D) $^{+4}$PD:DP :: 28:23 Glu (E) $^{+9}$PE:EP :: 32:17	Gly (G) $^{-9}$PG:GP :: 44:25	U C A G



a. 3 letter abbreviation

b. 1 letter symbol

c. % deviation from non-preference value (+ or -) in the choice of Proline (P) as neighbour

d. preference for placement of Proline (P) (left/right) in dipeptides

is reflected in CHART.C.I.12.

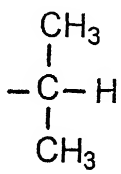
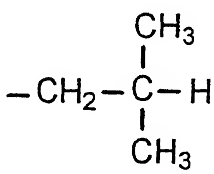
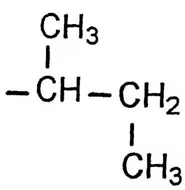
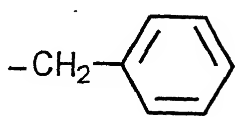
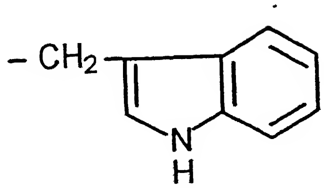
CHART.C.I.12 presents data base proline (P) dipeptide "Left-Right Preference" of amino acids having side chains without functional groups. It may be noted that an exception here is tryptophan. A superficial analysis would perhaps invoke steric factors associated with the coded amino acid side chains in the observed switch in preference profile from glycine (36-64) to phenylalanine (77-23). In any event it was felt that analysis of experimental results here would be less hazardous.

At this point it would be of interest to analyze the possible origins of selectivity with respect to peptide bond formation involving a central residue either from amino end (left preference) or the carboxyl end (right preference). As an illustration, if unprotected amino acids A and B are allowed to react and form peptide bond in water, using a soluble condensing agent, the extent of formation of 4 peptide bonds would be governed by, *inter alia*, the propensity of active ester formation, which, in the zwitter ionic profile of amino acid, would be related to the pKa profile of the carboxyl as well as amino group and the amidation step which would be more directly linked to the basicity of the amino group. Superimposed on this are steric factors associated with the active ester formation as well its transformation to the dipeptide. The steric considerations would play an important role in the second step, since, in this step, the two amino acid residues that are to form the peptide bond will have to come in close proximity.

CHART.C.I.13 outlines key steps in the formation of Pro-AA or Pro-Pro peptide bond. The possibilities shown here pertains to formation of Pro-AA and Pro-Pro when equivalent amounts of proline and another amino acid are allowed to form the peptide bond. In water, at pH 7.1, proline is almost exclusively in the zwitter ionic form ($k_{-1} >>> k_1$). Therefore the activated ester formation involving this residue must be via interaction of the zwitter ionic form of proline with the carbodiimide condensing agent. The activated ester formation here can be best rationalized on the basis of proton transfer to the carbodiimide from a zwitter ionic proline unit which is promoted by the addition

CHART.C.I.12

Data base: Proline (P) dipeptide "Left-Right" preference of amino acids having side chains without functional groups (AA).

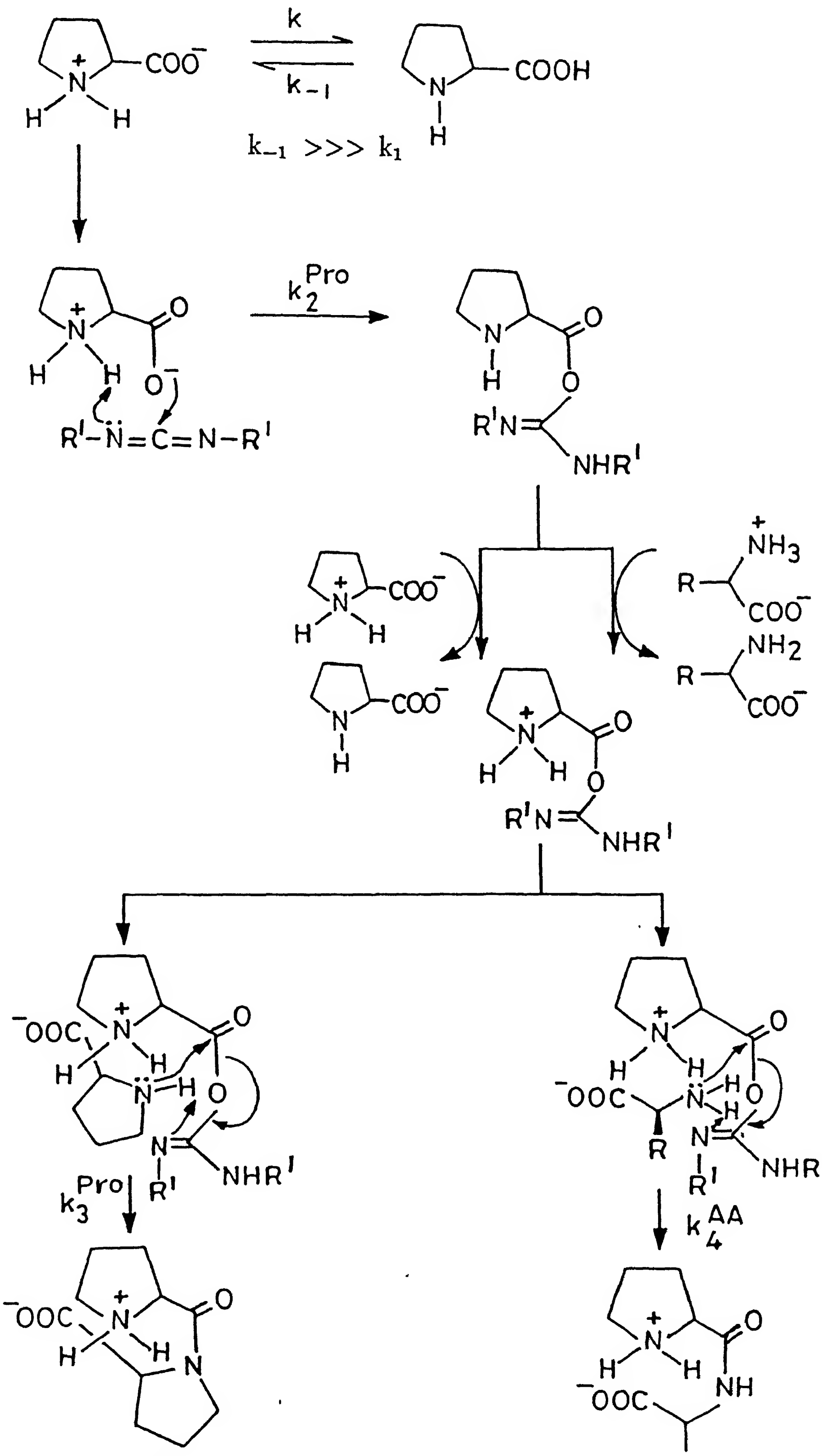
AMINO ACID	SIDE CHAIN	PROLINE DIPEPTIDES TOTAL	AA-P	:	P-AA	(as %)
Glycine	0	69	25	:	44	(36:64)
Alanine	$-CH_2-H$	65	32	:	33	(50:50)
Valine		55	22	:	33	(40:60)
Leucine		65	42	:	23	(65:35)
Isoleucine		45	32	:	13	(71:29)
Phenylalanine		30	23	:	07	(77:23)
Tryptophan		11	02	:	09	(18:82)

of the carboxyl end to the central carbon of the carbodiimide. In CHART.C.I.13 this aspect is depicted as taking place via a cyclic transition state. With this mechanistic profile it could be argued that the rigid proline frame would be of entropic advantages. It should be noted that the activated ester would also contain a highly basic proline NH in addition to the isourea unit. Thus, this represents a system which has in it a highly basic nitrogen function and a highly activated ester. These factors would make such a representation highly unstable in view of a great propensity for polymerization. Fortunately since the concentration of such an activated ester would be considerably smaller than the starting amino acid proline, as well as the partner AA, the protonation equilibria would dictate that almost all of the activated intermediate would be protonated at the pyrrolidine nitrogen, thus precluding polymerization and at the same time paving way for peptide bond formation. The peptidation step involving the protonated proline activated ester with either proline or a neighbour amino acid is illustrated in CHART.C.I.13. As could be seen here, the profile of the transition state would be largely influenced by the electrostatic interaction involving the protonated proline unit and the carboxylate end of either the proline or the amino acid residue. In this representation the steric factors leading to Pro-Pro peptide bond formation appears to be quite unfavourable compared to that involving a neighbouring amino acid. This could well account for the large deviation from the non preference value (-28 %) with reference to Pro-Pro in the basic set (CHART.C.I.11). The peptidation step itself is envisaged as taking place via a six membered transition state and is in profile very similar to those envisaged in normal peptide bond forming reactions.

A complementary picture is provided in CHART.C.I.14 with focus on activated ester formation involving the amino acid partner (AA). Hereagain a cyclic transition state pertaining to the formation of activated ester is envisaged from a zwitter ionic representation of the amino acid. In line with arguments presented earlier, concentration factors would largely dictate that the activated amino acid here would be largely protonated

CHART.C.I.13

Major controls in Pro-AA / Pro-Pro peptide bond formation mediated by water soluble carbodiimide in water (pH 7.1).

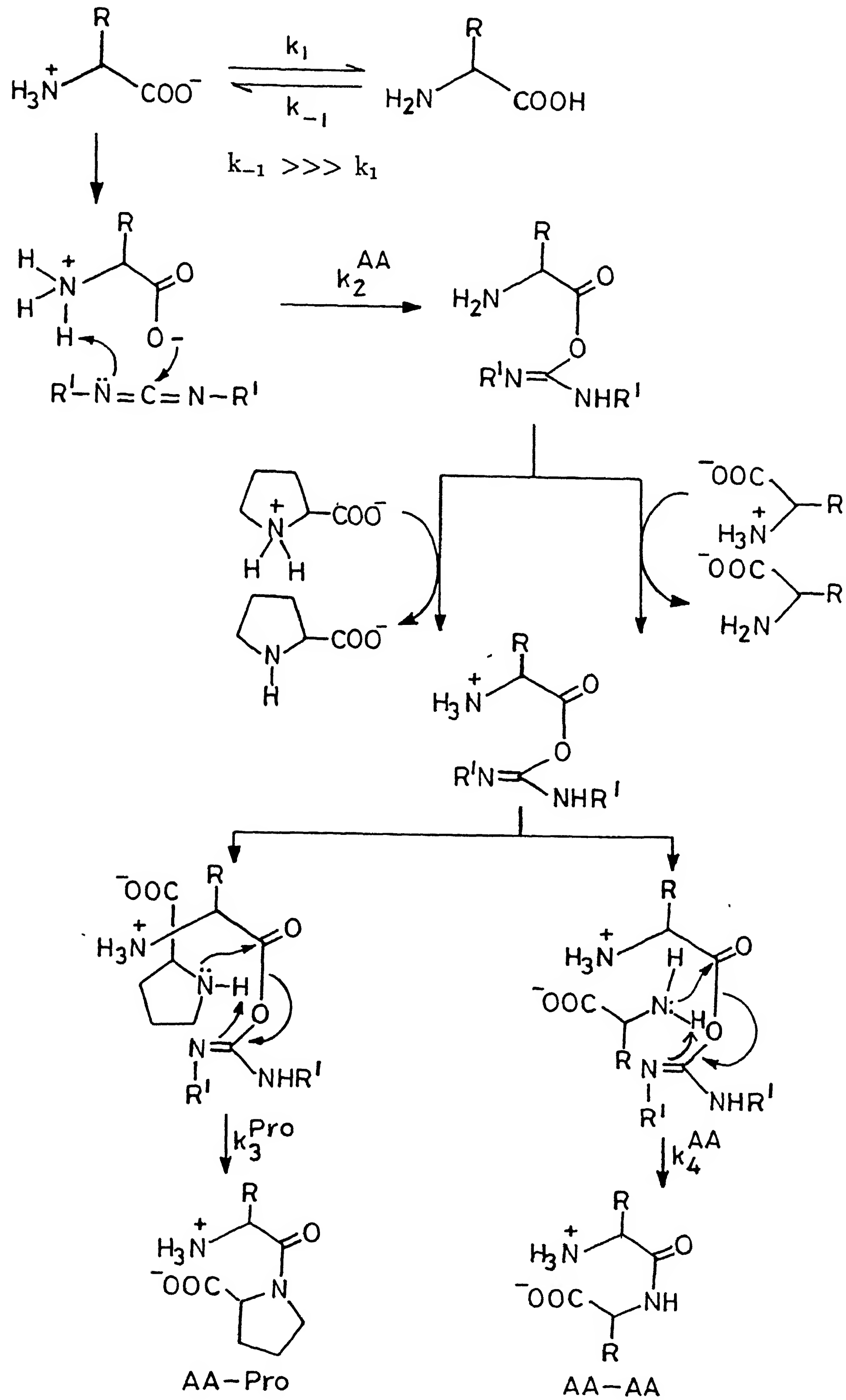


Pro-Pro

Pro-AA

CHART.C.I.14

Major controls in AA-Pro / AA-AA peptide bond formation mediated by water soluble carbodiimide in water (pH 7.1).



at the amino terminal end as envisaged in CHART.C.I.14. The subsequent peptidation step would also be largely dictated by electrostatic interactions and the reaction itself is best envisaged as proceeding via a six membered transition state. It is obvious from CHART.C.I.14 that the propensity for formation of either AA-Pro peptide bond or AA-AA peptide bond would be dictated by the magnitude of the interaction involving that of proline and the side chain of the α-amino acid on one hand and that arising from amino acid side chain interactions on the other (CHART.C.I.14).

In the delineation of neighbour preferences, the observed non preference index from the basic set could be either due to the intrinsic properties of the side chains or due to extrinsic controls that could modify the structure and reactivity profile of the amino acids. The easiest way to bring about such a modification would be by blocking either of the N or C terminal ends. The analysis of peptide bond formation using Ts-Pro / Pro-OMe on the one hand and Ts-AA /AA-OMe on the other, instead of Pro and AA (CHART.C.I.13 and CHART.C.I.14) would be of interest. Such an analysis is illustrated in CHART.C.I.15 and CHART.C.I.16.

Pathways involved in the peptidation of Ts-Pro with either Pro-OMe or AA-OMe is shown in CHART.C.I.15. These are similar to that involved in peptidation of protected amino acid substrates. The formation of the activated ester from Ts-Pro can best be explained on the basis of a six membered transition state, initiated by carboxyl proton abstraction by the diimide condensing agent. A comparison of CHART.C.I.13 and CHART.C.I.15 would immediately show that in the peptidation step involving either the Pro-OMe or AA-OMe, although the mechanistic pathway is identical in terms of the putative six membered transition state, the lack of electrostatic controls would make the course of the reaction almost directed by steric differentiation of the pyrrolidine in relation to the side chain of coded amino acids. On the basis of this rationale it could be predicted that the preference here would be for Ts-Pro-AA-OMe formation rather than Ts-Pro-Pro-OMe.

CHART.C.I.15

Major controls in Ts-Pro-AA-OMe / Ts-Pro-Pro-OMe peptide bond formation mediated by water soluble carbodiimide in water (pH 7.1).

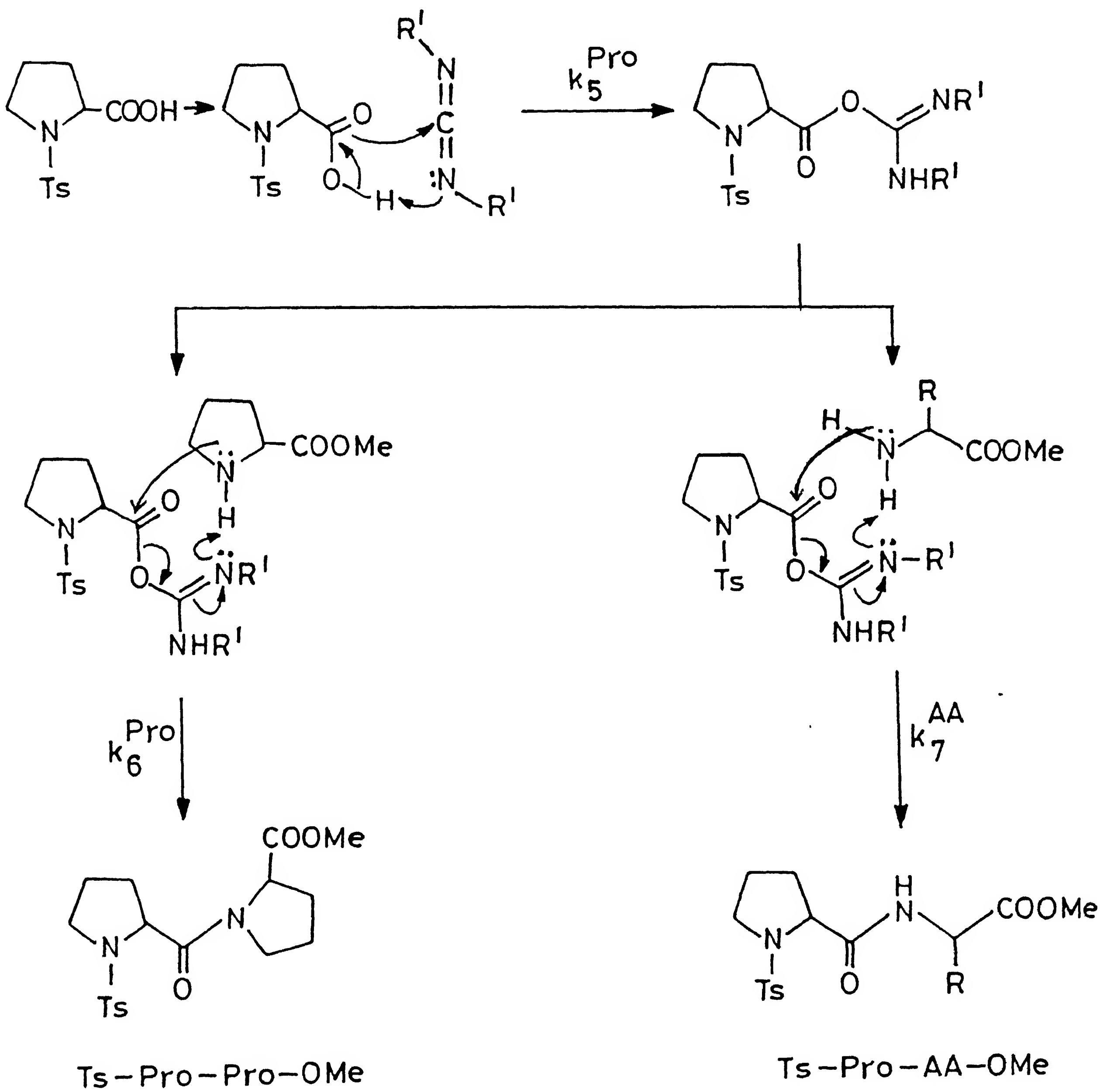
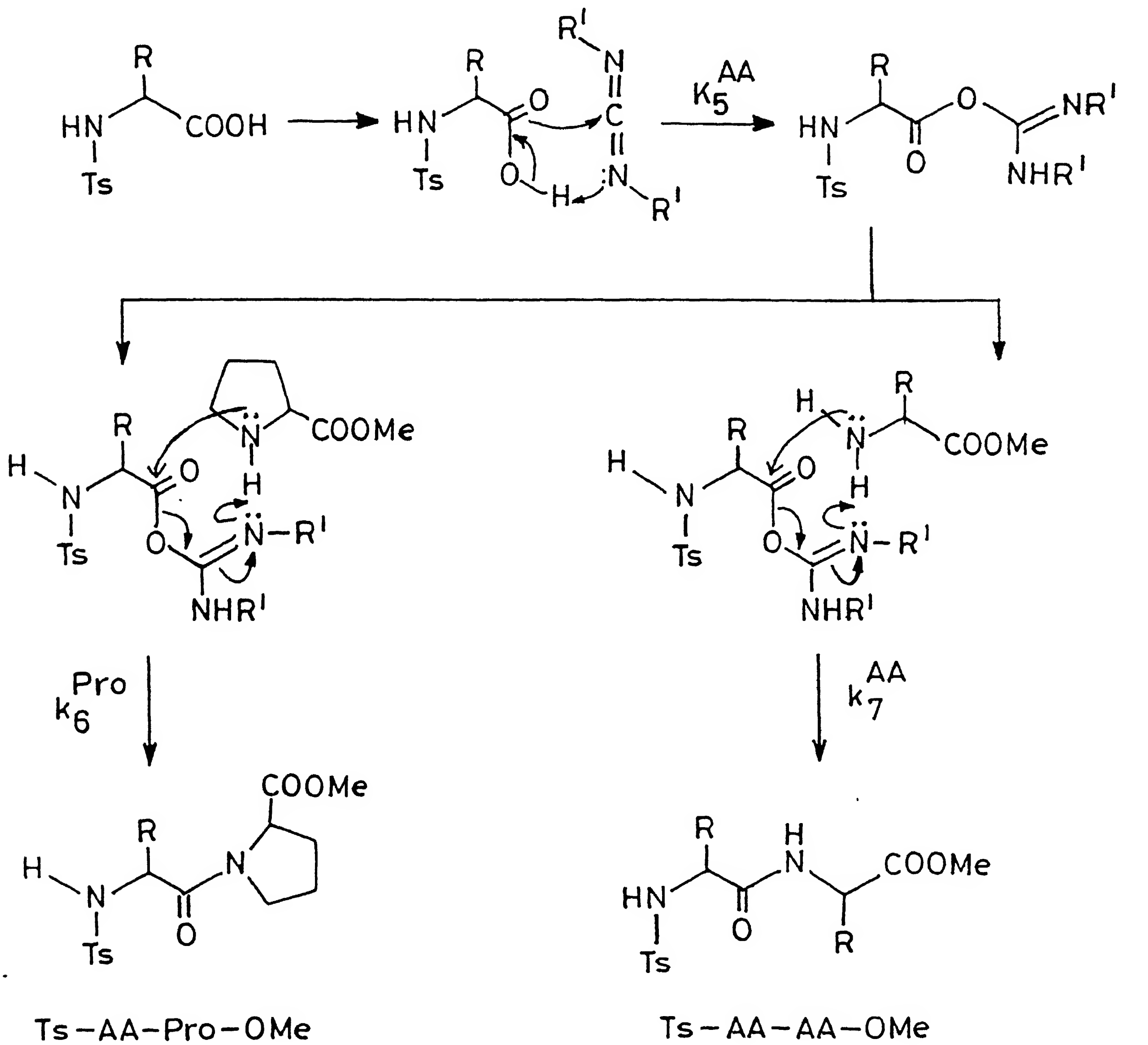


CHART.C.I.16

Major controls in Ts-AA-Pro-OMe / Ts-AA-AA-OMe peptide bond formation mediated by water soluble carbodiimide in water (pH 7.1).



In CHART.C.I.16 are shown similar controls that would operate in the peptidation of Ts-AA with either Pro-OMe or AA-OMe. Here also the formation of the activated ester proceeds by a similar mechanism. In the subsequent peptidation also the factors involved here are quite similar leading to prediction that an activated Ts-AA is more likely to form peptide bond with AA-OMe in preference to Pro-OMe.

The different mechanistic profiles envisaged pertaining to the formation of peptide bonds involving free amino acids and protected substrates, which have not received attention, are noteworthy. Although CHARTS.C.I.13 to C.I.16 provide an opportunity to make superficial conclusions pertaining to peptidation, in view of various other factors that may influence the course of the reaction, experimental endeavours would be the only way to gain further insight relating to the genesis of preferences established for such processes.

The experimental focus of the present study was to correlate the actual propensity for preferential peptidation involving proline and that observed from the basic data set with α-amino acids having side chains without functional groups. Thus in addition to proline, the amino acids selected for experimental study were glycine, leucine, phenylalanine and tryptophan. Simply stated, the experiment comprises protocol for the reaction of proline with each of these amino acids, either free or protected followed by analysis of dipeptide profile based on authentic standards, prepared separately.

As the first step, 13 dipeptides, falling into 4 sets were prepared and experimental conditions were delineated for their clear cut differentiation in HPLC. The set of dipeptides and their HPLC profiles are presented in TABLE.C.I.2.

Peptide bond formation with proline as the core residue was examined with glycine, leucine, phenylalanine and tryptophan in water at pH 7.1 using the water soluble carbodiimide, 1-Cyclohexyl-3-(2-morpholinoethyl) carbodiimide metho-p-toluene sulfonate (WSCDI).

The experimental protocol in each of the 4 sets can be generalized as follows: clear

TABLE.C.I.2

HPLC PROFILE OF AUTHENTIC N,C-PROTECTED DIPEPTIDES

Subset	Dipeptide	Mobile Phase	Flow Rate	Retention Time (Min)
	Ts-Pro-Pro-OMe			8.85
A	Ts-Pro-Gly-OMe	MeOH : H_2O	0.8 mL/min	6.86
	Ts-Gly-Pro-OMe	::		6.39
	Ts-Gly-Gly-OMe	60 : 40		5.25
	Ts-Pro-Pro-OMe			6.28
B	Ts-Pro-Leu-OMe	MeCN : H_2O	0.8 mL/min	10.55
	Ts-Leu-Pro-OMe	::		9.29
	Ts-Leu-Leu-OMe	60 : 40		12.58
	Ts-Pro-Pro-OMe			6.26
C	Ts-Pro-Phe-OMe	MeCN : H_2O	0.8 mL/min	11.16
	Ts-Phe-Pro-OMe	::		9.25
	Ts-Phe-Phe-OMe	60 : 40		13.67
	Ts-Pro-Pro-OMe			6.24
D	Ts-Pro-Trp-OMe	MeCN : H_2O	0.8 mL/min	9.3
	Ts-Trp-Pro-OMe	::		8.0
	Ts-Trp-Trp-OMe	60 : 40		10.66

aqueous solution of equivalent amounts of Pro and the target amino acid (Gly / Leu / Phe / Trp) with 3 equivalents of the WSCDI in water (5 ml per mmol of amino acid) were left stirred at room temperature for 48 hours. There was practically no change in the pH of the medium during the course of the reaction. The reaction mixture was treated with tosyl chloride - 2N NaOH, left stirred for 4 hours, filtered, ice-cooled, adjusted to pH 2 with 5N HCl, saturated with NaCl, extracted with ethylacetate, dried over anhydrous $MgSO_4$, evaporated, dissolved in minimum amount of methanol, cooled, treated with ethereal CH_2N_2 and evaporated. The dipeptide distribution amongst the 4 possibilities was determined by HPLC by comparison with authentic dipeptide samples. Duplicate runs were made in each case.

Thus FIGURES.C.I.1a, C.I.1b, C.I.1c and C.I.1d illustrate the HPLC profile of authentic samples belonging to each set involving, respectively, Gly, Leu, Phe and Trp. The actual experimentally determined dipeptide composition as HPLC profiles are presented in FIGURES.C.I.2a, C.I.2b, C.I.2c and C.I.2d pertaining to respectively Gly, Leu, Phe and Trp.

The observed dipeptide distribution from these sets of experiment is presented in TABLE.C.I.3. Along with the percentage distribution of dipeptides are presented within brackets the estimated yields by HPLC of dipeptide present in the mixture. The complete absence of Pro-Pro is an expectation with the envisaged mechanistic pathways and a highly unfavourable non preference index of -28% for this unit (CHART.C.I.11). A surprising picture here is the lack of formation of AA-AA peptide bond formation (AA = Gly, Leu, Phe, Trp), particularly, since similar experiments performed with N or C protected ends afforded substantial percentages of such homo peptides (*vide infra*). A possible rationalization could be that in the case of Gly, because its involvement is seen only as Pro-Gly (100%), any activated ester form involving this residue might rapidly polymerize. With reference to examples involving Leu, Phe and Trp, a possible explanation can be seen from the mechanistic pathways envisaged in CHART.C.I.14. As

FIGURE.C.I.1a

HPLC profile of authentic dipeptides for Proline and Glycine

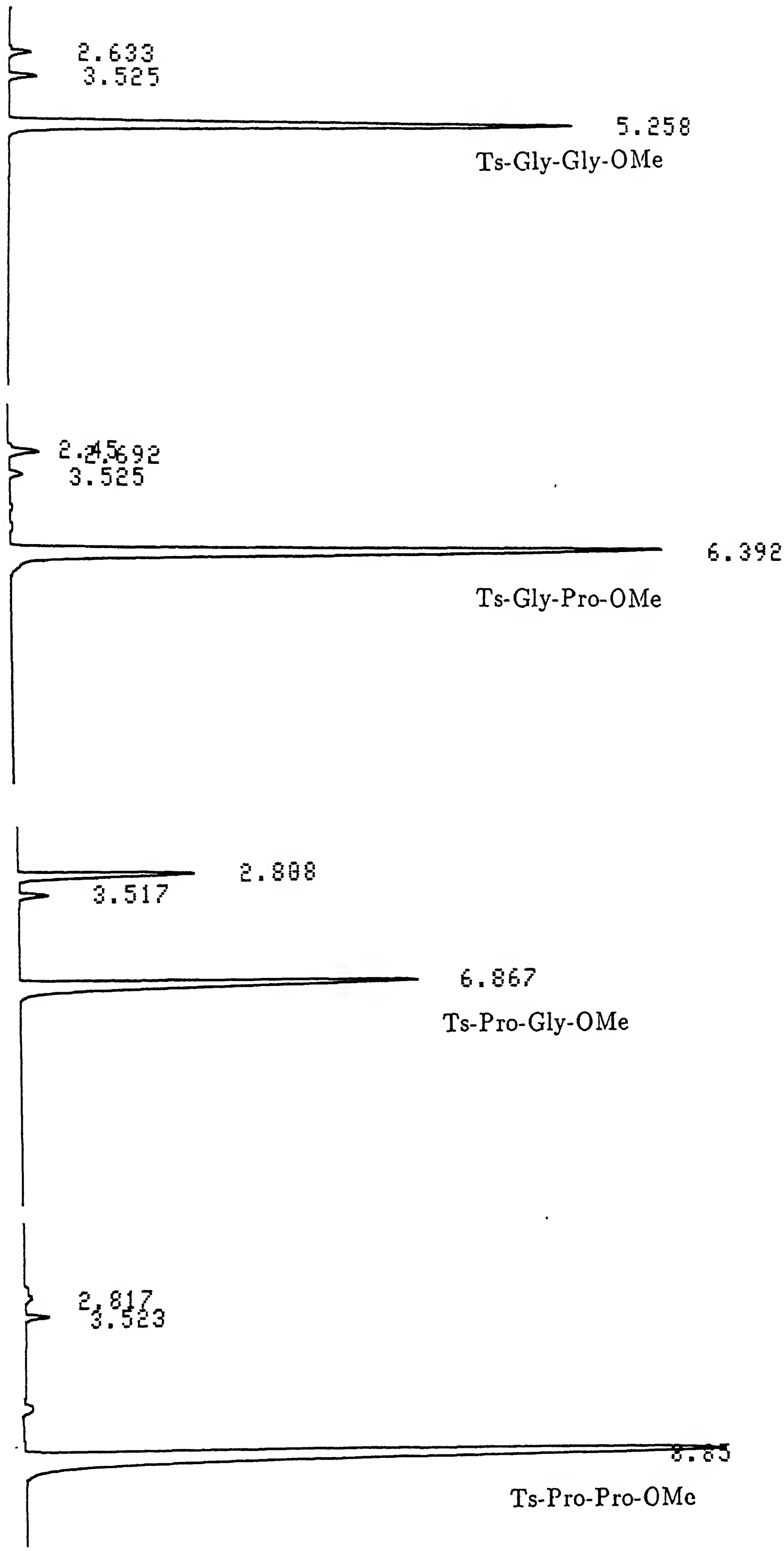


FIGURE.C.I.2a

Dipeptide distribution in the condensation of Proline with Glycine in presence of WSCDI in water

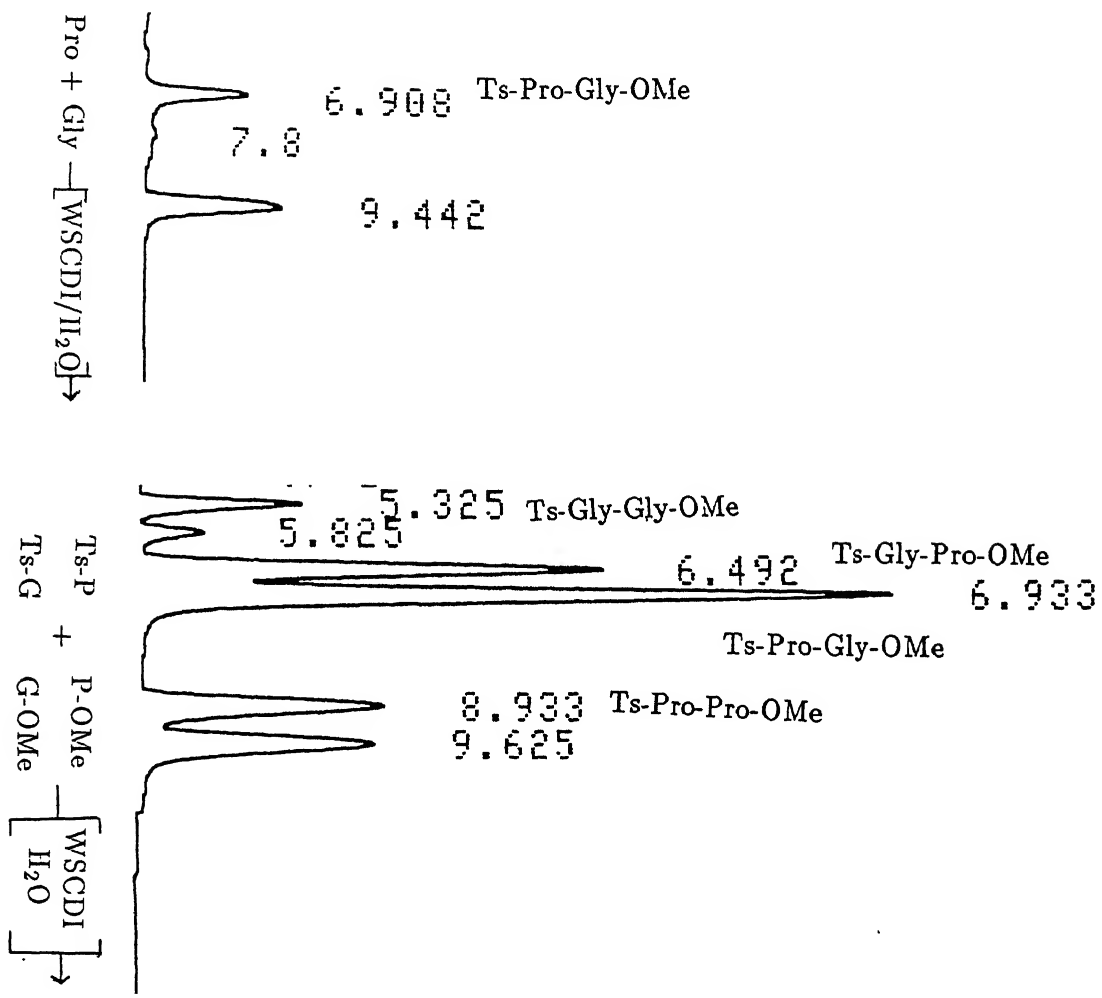


FIGURE.C.I.1b

HPLC profile of authentic dipeptides for Proline and Leucine

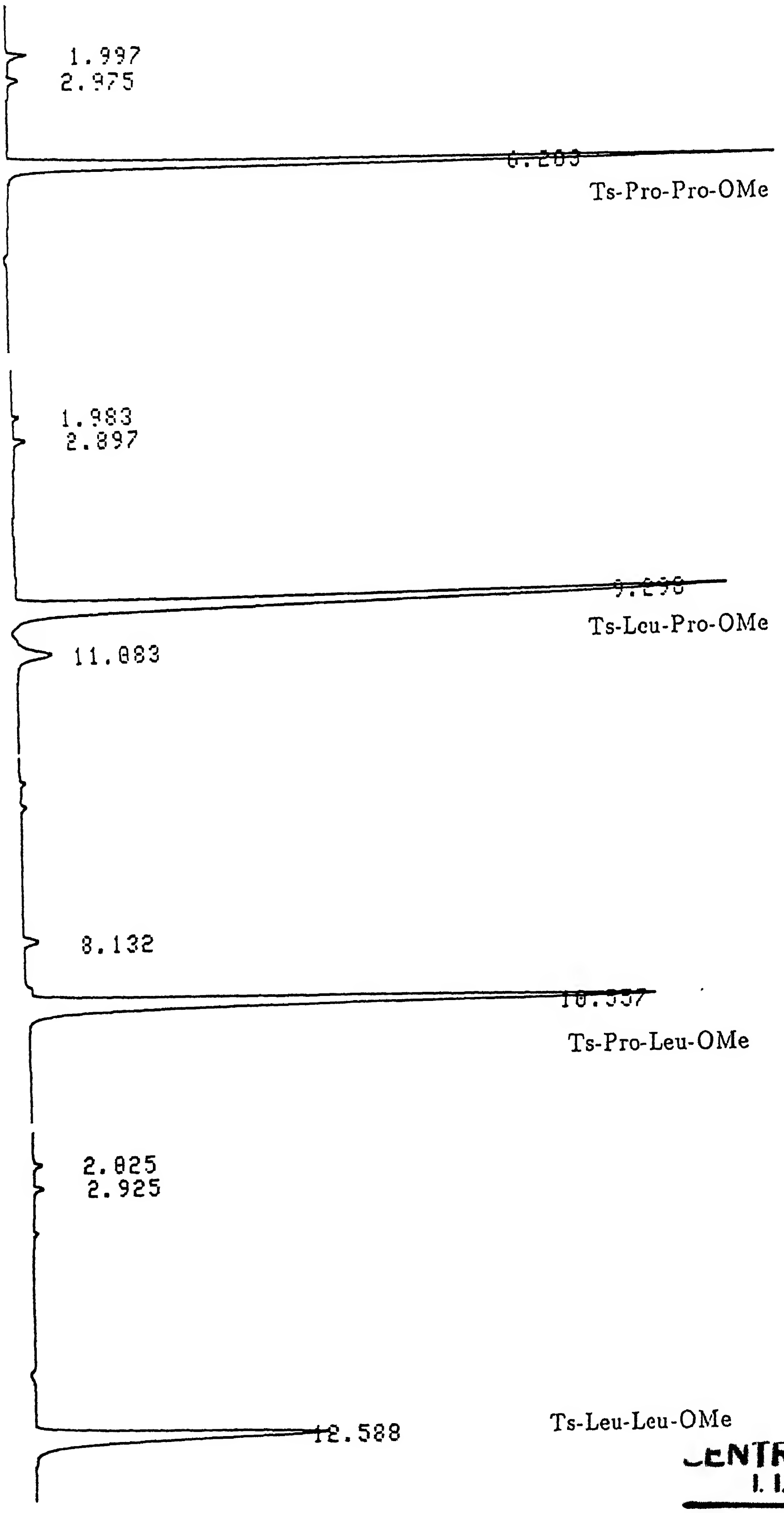


FIGURE.C.I.2b

Dipeptide distribution in the condensation of Proline with Leucine in presence of WSCDI in water or water-acetonitrile mixture

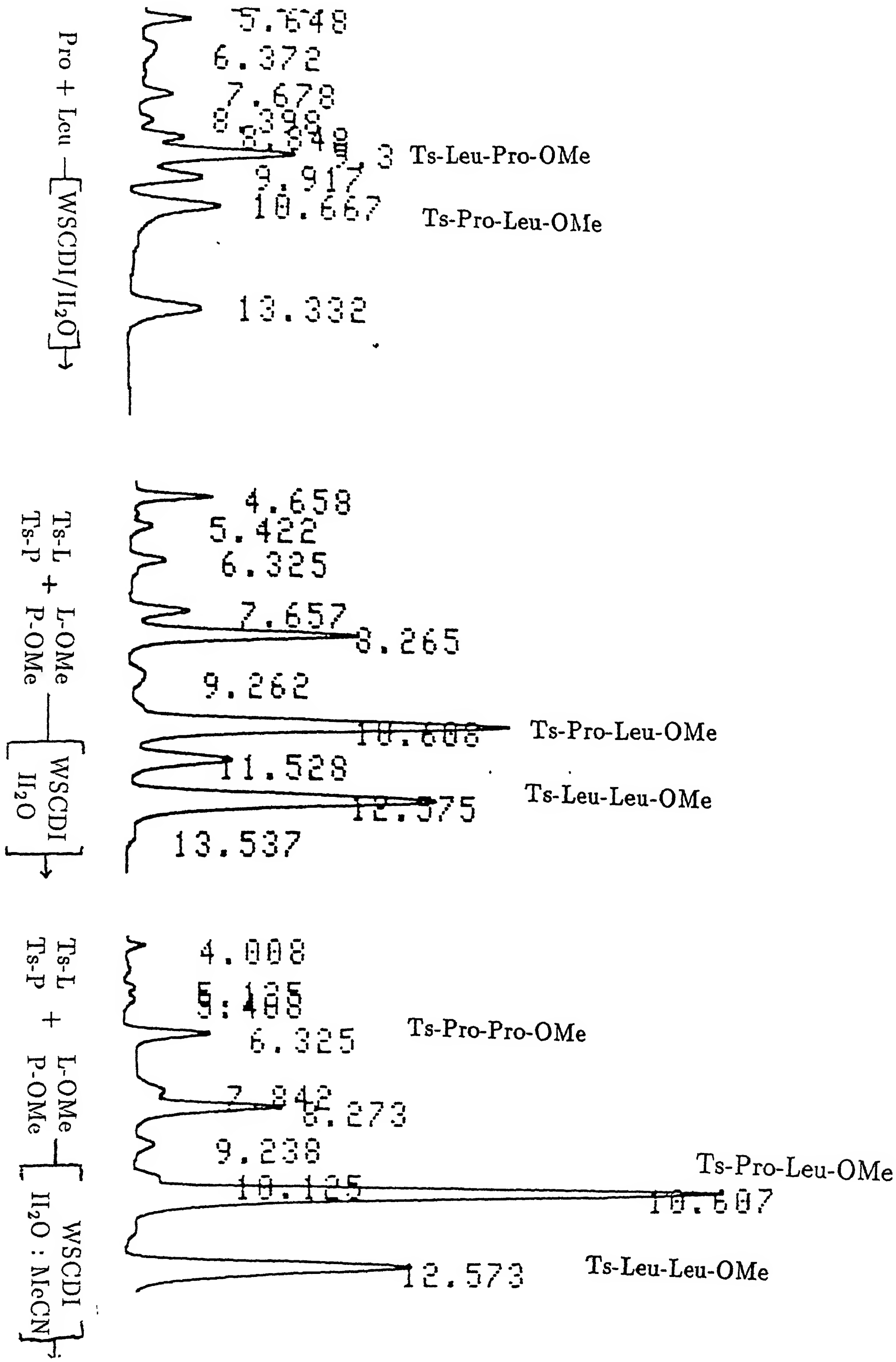


FIGURE.C.I.1c
IIPLC profile of authentic dipeptides for Proline and Phenylalanine

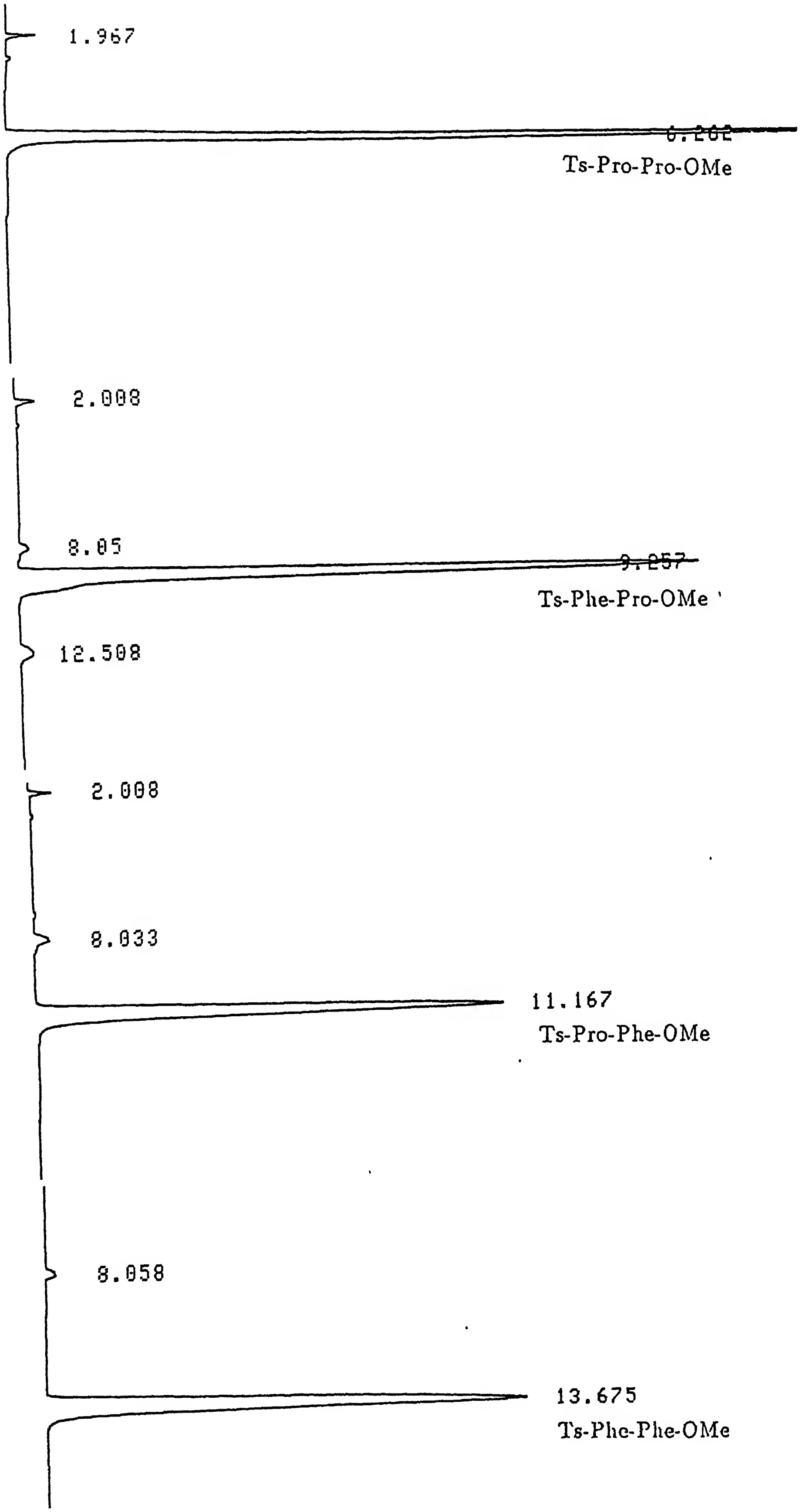


FIGURE.C.I.2c

Dipeptide distribution in the condensation of Proline with Phenylalanine in presence of WSCDI in water or water-acetonitrile mixture

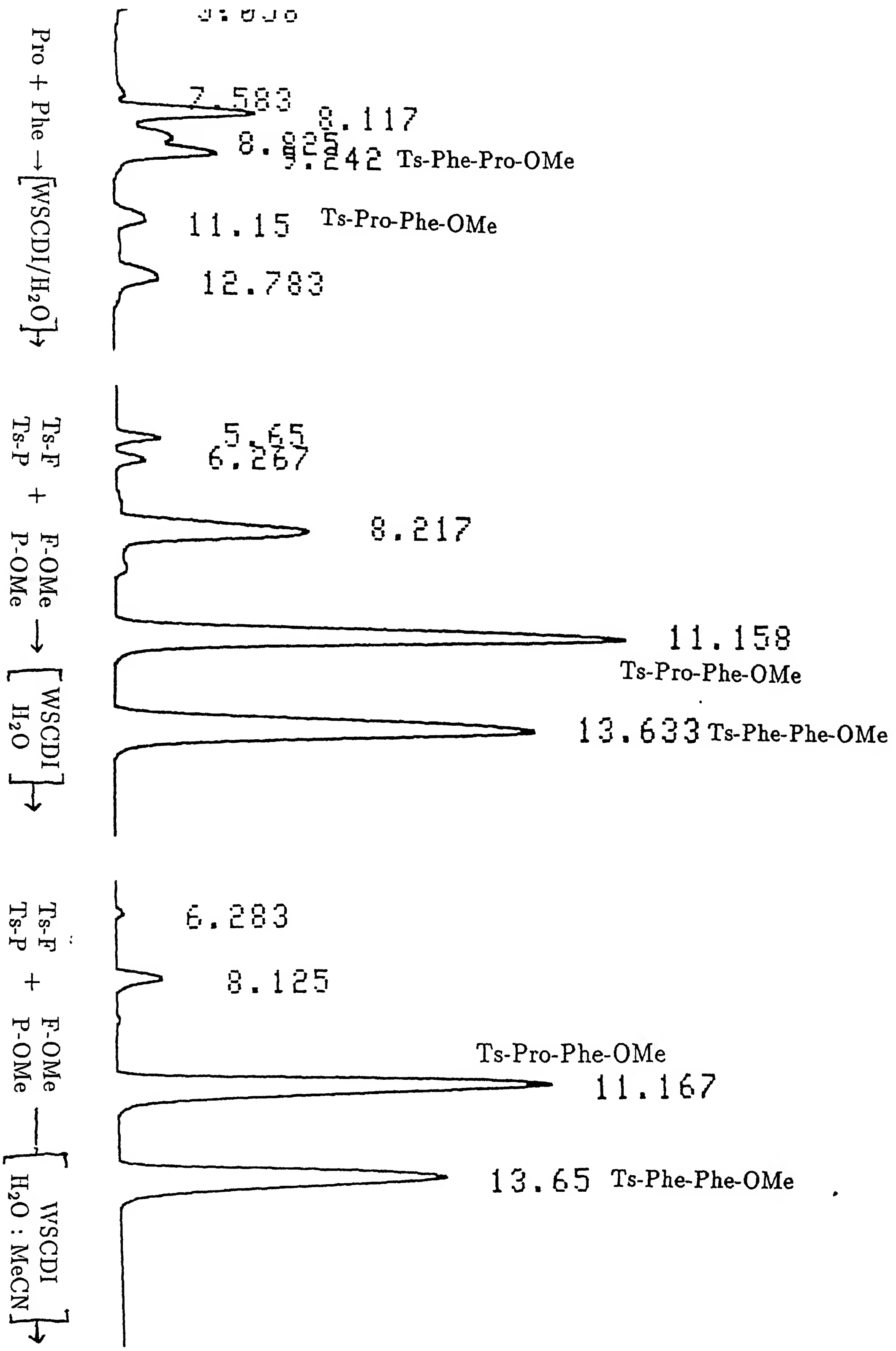


FIGURE.C.I.1d

HPLC profile of authentic dipeptides for Proline and Tryptophan

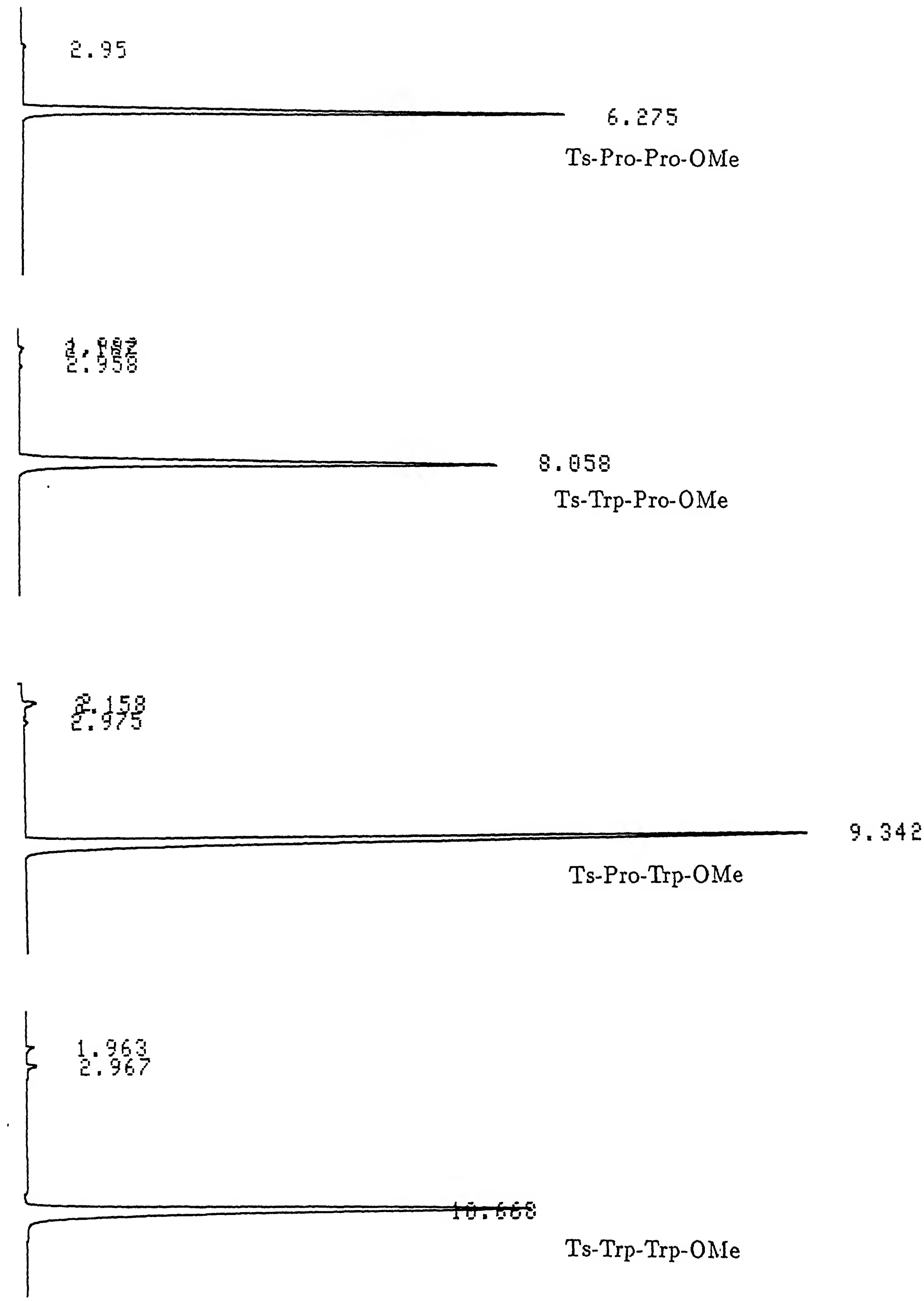
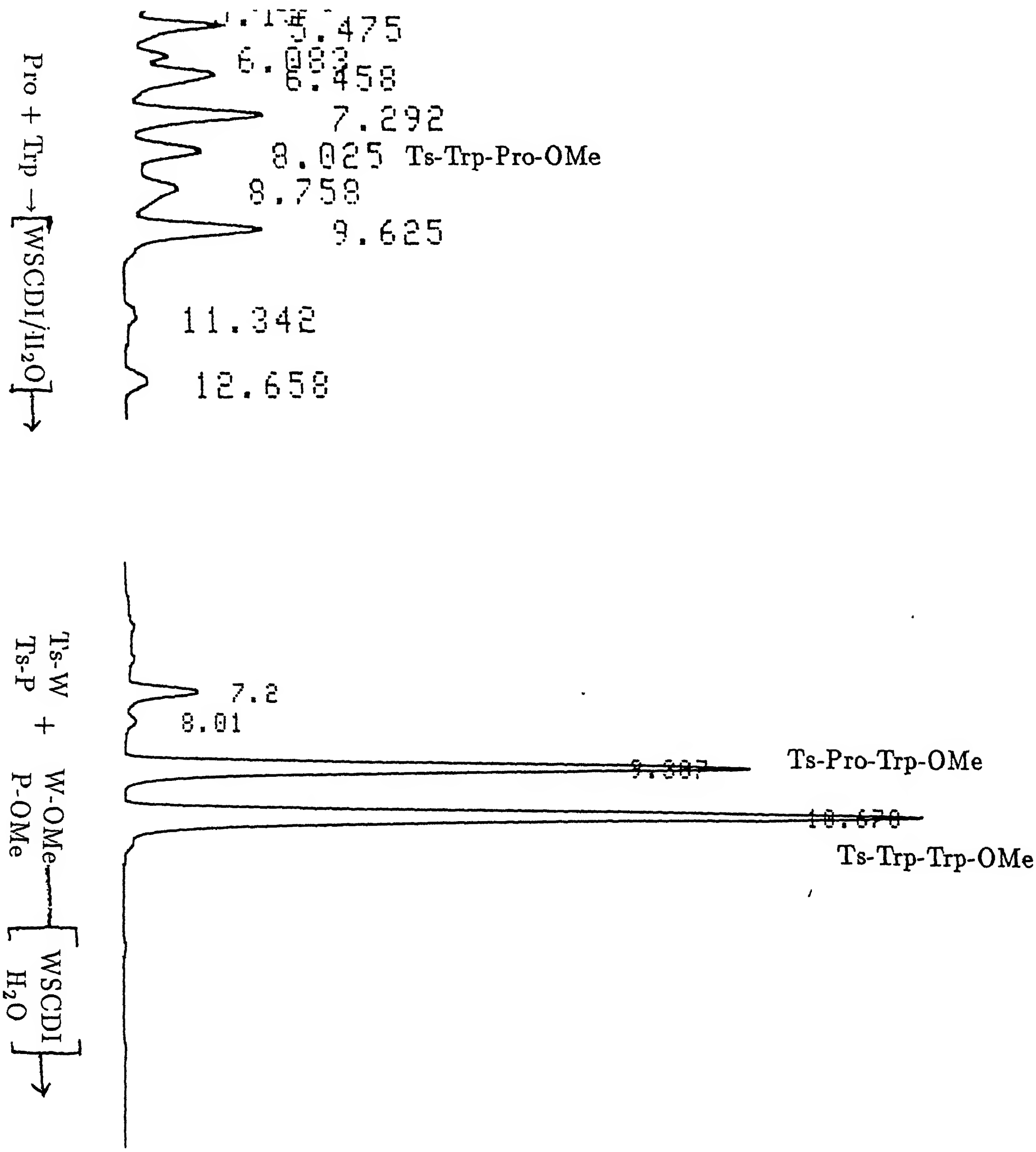


FIGURE.C.I.2d

Dipeptide distribution in the condensation of Proline with Tryptophan in presence of WSCDI in water

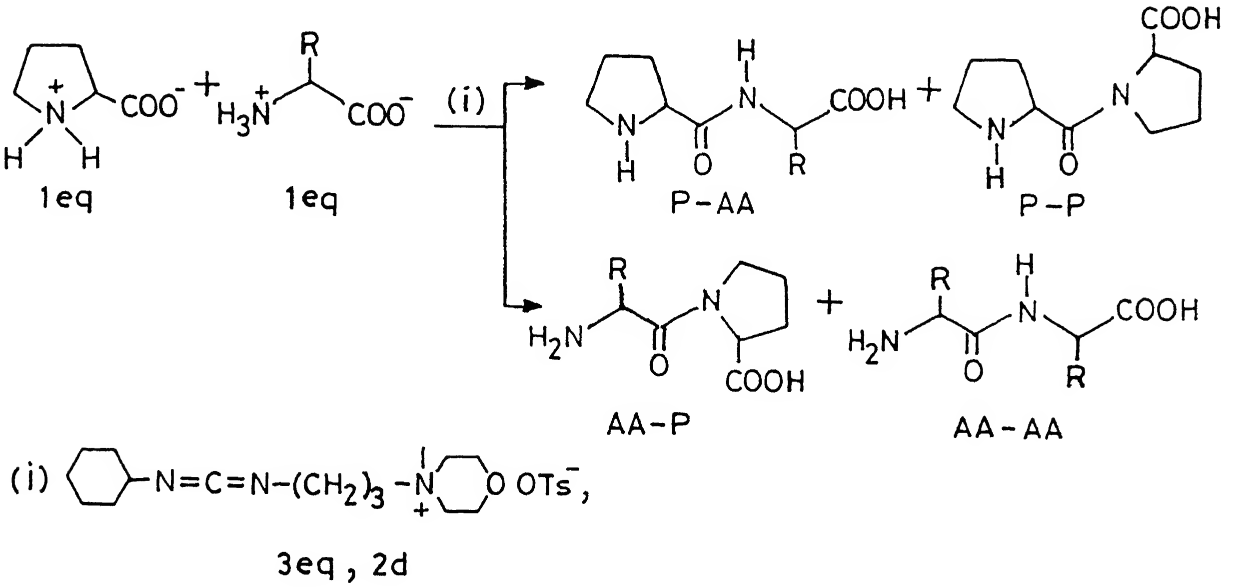


expected from this, when the side chain is sterically demanding, the activated amino acid preferentially interacts with proline residue. A pointer in this direction is that, as could be seen from TABLE.C.I.3, whilst the percentages of AA-Pro increases with increasing steric demands of the α-amino acid side chains, the percentages of Pro-AA steadily decreases from 100 to 0%, the latter would reflect the fact that the activated proline would become decreasingly available in dipeptide formation with increasing steric requirements of the amino acid side chains.

The experimentally determined left-right preferences involving dipeptide with Pro and Gly, Leu, Phe and Trp are presented in CHART.C.I.17, which show not only a clear cut correlation between the nature of the side chain of the partner amino acid but also that this preference profile, except for Trp, is in complete agreement with left-right preference derived from data set (CHART.C.I.12). Significant is the fact that whilst CHART.C.I.12 shows a consistent increase in preference with respect to the placement of the amino acid left to Pro on increasing the steric demands of side chains which are devoid of functional groups, an aspect which is in complete agreement with experimental results (CHART.C.I.17), a notable exception is Trp which whilst in our experiments show the expected trend, the data base provides an opposite profile. This can arise from two reasons : the obvious is, that the rather low (11) number of Pro-Trp dipeptides in the set preclude a viable interpretation, an interesting and quite reasonable alternative is that the Trp residues in the basic set arose during evolution by mutation. This view is supported by the fact that this amino acid is generally considered as late and perhaps only entrant to the code complement. Both experiment and theory tend to support the notion that carbon substituents placed at the γ position of α-amino acid side chains tend to shift, progressively the preference profile from Pro-AA (right) to AA-Pro (left). It seems logical to conclude that where the experimental sets are enlarged to include residues such as alanine, valine and isoleucine, the trend would be predictable and along those seen in CHART.C.I.12 and CHART.C.I.17. From the experimental vantage such

TABLE.C.I.3

Dipeptide distribution in the condensation of Proline (P) and amino acids lacking functional groups (AA), mediated by water soluble carbodiimide in water (pH 7.1).



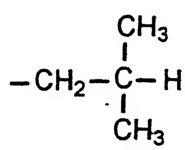
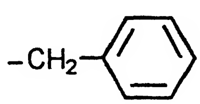
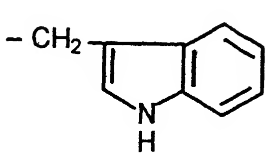
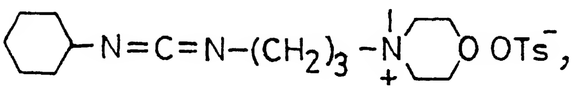
AMINO ACID (AA)	R	P-P	AA-P	P-AA	AA-AA
Glycine	H	0	0	100 (7.3%)	0
Leucine		0	59 (16%)	41 (11%)	0
Phenylalanine		0	74 (12.7%)	26 (4.4%)	0
Tryptophan		0	100 (7.6%)	0	0

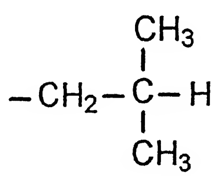
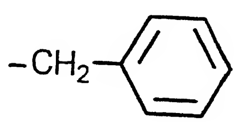
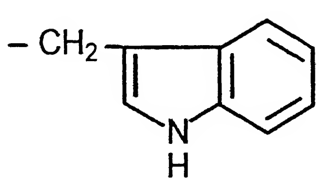
CHART.C.I.17

Experimentally determined Proline (P) dipeptide "Left-Right" preference of amino acids (AA) having side chains without functional groups in water.

$$\underset{1eq}{P} + \underset{1eq}{AA} \xrightarrow{(i)} \text{P-AA} + \text{P-P} + \text{AA-P} + \text{AA-AA}$$

(i) 

3eq , 2d

AMINO ACID (AA)	SIDE CHAIN	AA-P	:	P-AA	(as %)
Glycine	0	0	:	7.3	(0:100)
Leucine		16	:	11	(59:41)
Phenylalanine		12.7	:	4.4	(74:26)
Tryptophan		7.6	:	0	(100:0)

set of experiments are presented in FIGURES.C.I.2a, C.I.2b, C.I.2c and C.I.2d and in case of Leu and Phe where the reactions were also performed in water:acetonitrile in FIGURES. C.I.2b and C.I.2c.

The dipeptide distribution profile in the condensation N/C terminal blocked Pro and N/C terminal blocked Gly, Leu, Phe and Trp is summarized in TABLE.C.I.4, along with the percentage distribution is indicated within brackets the actual percentages of yield of dipeptides present in the reaction mixture. Additionally, the results in case of Leu and Phe, where the reaction was conducted in water:acetonitrile are also presented here.

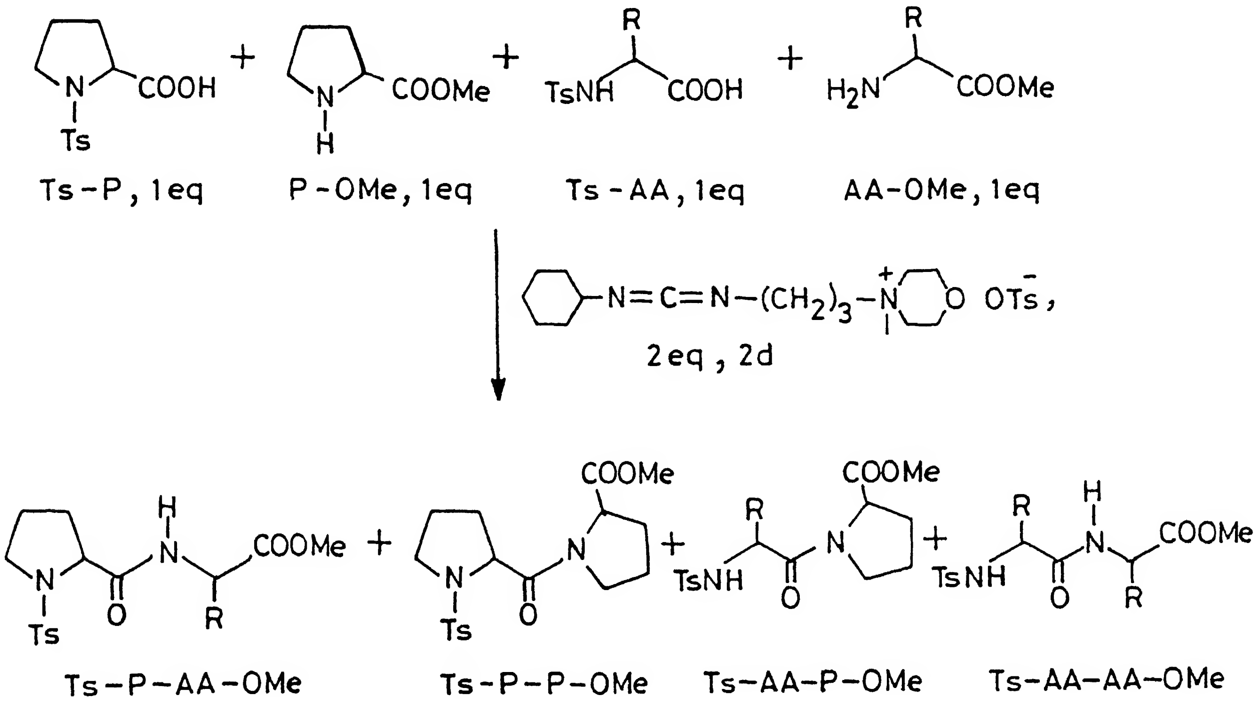
The most significant observation arising from this study as could be readily seen from TABLE.C.I.4 is that regardless of their structures, AA-OMe residue here is extremely effective in peptide bond formation. Indeed, this table reflects, almost exclusively, the competition of activated Ts-Pro and Ts-AA with AA-OMe in peptide bond formation. It could also be seen that, by and large these two processes are equally effective, no left-right preference is seen here. In this set is also reflected the difficulty in the formation of the Pro-Pro peptide bond. TABLE.C.I.4 also shows that in changing from water to water:acetonitrile (4:1) does not affect the outcome of the preferences.

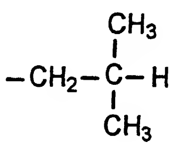
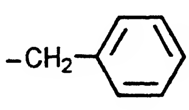
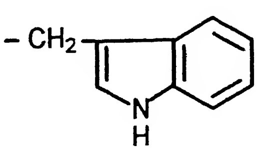
As stated previously, whereas electrostatic and steric factors are involved in the peptidation involving free amino acids (CHART.C.I.13 and CHART.C.I.14), in the case of N/C blocked substrates the peptidation is controlled largely by the steric factors and consequently the transition state here is more relaxed. Thus it is quite obvious from CHART.C.I.15 and CHART.C.I.16 that peptidation involving AA-OMe would be more preferred over that with Pro-OMe, as has been found to be the case.

In CHART.C.I.18 and CHART.C.I.19 are presented experimentally determined N/C terminal blocked Pro dipeptide preference for N/C terminal blocked Gly, Leu, Phe and Trp in water and water:acetonitrile (4:1) respectively. It could be seen from here that, in sharp contrast to the parallel set with unprotected set CHART.C.I.17, no left-right preference are seen. Thus whilst preferences for Pro-AA and AA-Pro formation are

TABLE.C.I.4

Dipeptide distribution in the condensation of N or C terminal blocked Proline (P) and N or C terminal blocked amino acids lacking functional groups (AA), mediated by water soluble carbodiimide in water (pH 7.1) and in water : acetonitrile (4 : 1)*.



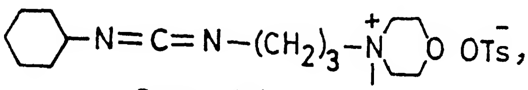
AMINO ACID	R	A	B	C	D
Glycine	H	12 (9.4%)	24 (19%)	57 (46%)	07 (5.3%)
Leucine		0, 5* (0%), (5%)	0, 0* (0%), (0%)	54, 62* (37%), (56%)	46, 33* (32%),(30%)
Phenylalanine		0, 0* (0%), (0%)	0, 0* (0%), (0%)	53, 56* (47%), (68%)	47, 44* (42%), (54%)
Tryptophan		0 (0%)	0 (0%)	45 (42%)	55 (51%)

A=Ts-P-P-OMe, B=Ts-AA-P-OMe, C=Ts-P-AA-OMe, D=Ts-AA-AA-OMe

CHART.C.I.18

Experimentally determined N or C terminal bolcked Proline (P) dipeptide preference for N or C terminal blocked amino acids (AA) having side chains without functional groups in water.

Ts-P +	P-OMe	$\xrightarrow{(i)}$	Ts-P-AA-OMe + Ts-P-P-OMe +
leq	leq		
Ts-AA +	AA-OMe		Ts-AA-P-OMe + Ts-AA-AA-OMe
leq	leq		

(i) 

2eq , 2d

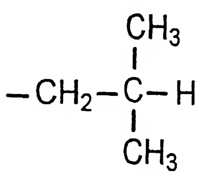
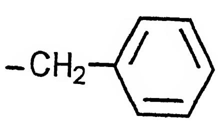
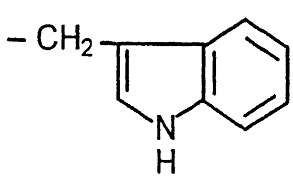
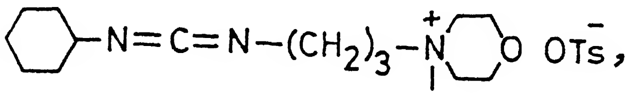
AMINO ACID	SIDE CHAIN	Ts-AA-P-OMe	:	Ts-P-AA-OMe	(as %)
Glycine	0	19	:	46	(29:71)
Leucine		0	:	37	(0:100)
Phenylalanine		0	:	47	(0:100)
Tryptophan		0	:	42	(0:100)

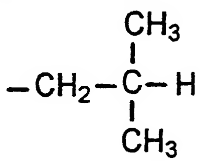
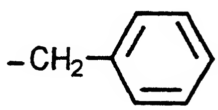
CHART.C.I.19

Experimentally determined N or C terminal bolcked Proline (P) dipeptide preference for N or C terminal blocked amino acids (AA) having side chains without functional groups in water : acetonitrile (4:1).

Ts-P + P-OMe $\xrightarrow{\text{(i)}}$ Ts-P-AA-OMe + Ts-P-P-OMe +
1eq 1eq
Ts-AA + AA-OMe Ts-AA-P-OMe + Ts-AA-AA-OMe
1eq 1eq

(i) 

2eq , 2d

AMINO ACID	SIDE CHAIN	Ts-AA-P-OMe	:	Ts-P-AA-OMe	(as %)
Leucine		0	:	56	(0:100)
Phenylalanine		0	:	68	(0:100)

modulated by the side chains of the amino acids in water that with N/C protected substrates do not exhibit any such differentiation.

These experiments provide support to the view that intrinsic factors can lead to the left-right preference observed in peptide bond formation involving proline and amino acids having side chains devoid of functional groups.

An important objective of the present study is to demonstrate that larger peptides could be prepared in a selective manner by blockwise condensation of smaller units. A demonstration of possibilities in this direction not only would strengthen the knowledge pertaining to preferences in peptide bond formation but also would enable the assembly of peptide segments with minimum of protection and number of steps involved. In a larger sense, any pointer in this direction would be an attestation to the belief that peptide of larger sequence can be produced in a sequence specific manner based on selectivities which control the formation of peptide bond. In the present work this aspect has been experimentally examined using N/C terminal protected dipeptide blocks Pro-Phe and Phe-Pro. Although such a series could have been carried out using unprotected dipeptides, protected precursors were considered more desirable for two reasons : The work carried out thus far has clearly showed neighbour preference selectivity amongst individual amino acids is markedly altered on N/C protection. Thus it can be concluded that the reactivity profile in the formation of tetrapeptide from a combination of blocks Pro-Phe and Phe-Pro can not be predicted on the basis of individual propensity of these residues. Second, from a more practical vantage, the experimental procedures are easier and the yields are higher when N/C protected blocks are used.

In CHART.C.I.20 is presented major controls that could direct peptidation from the vantage of a terminal Pro activation. In water mediated by WSCDI, the pathways envisaged here are precisely similar to that for Ts-Pro shown in CHART.C.I.15. A noticeable difference between CHART.C.I.15 and CHART.C.I.20 is the highly increased steric requirements of Pro end substituents. This factor can be logically expected to

CHART.C.I.20

Major controls in Boc-Phe-Pro-Pro-Phe-OMe / Boc-Phe-Pro-Phe-Pro-OMe peptide bond formation mediated by water soluble carbodiimide in water.

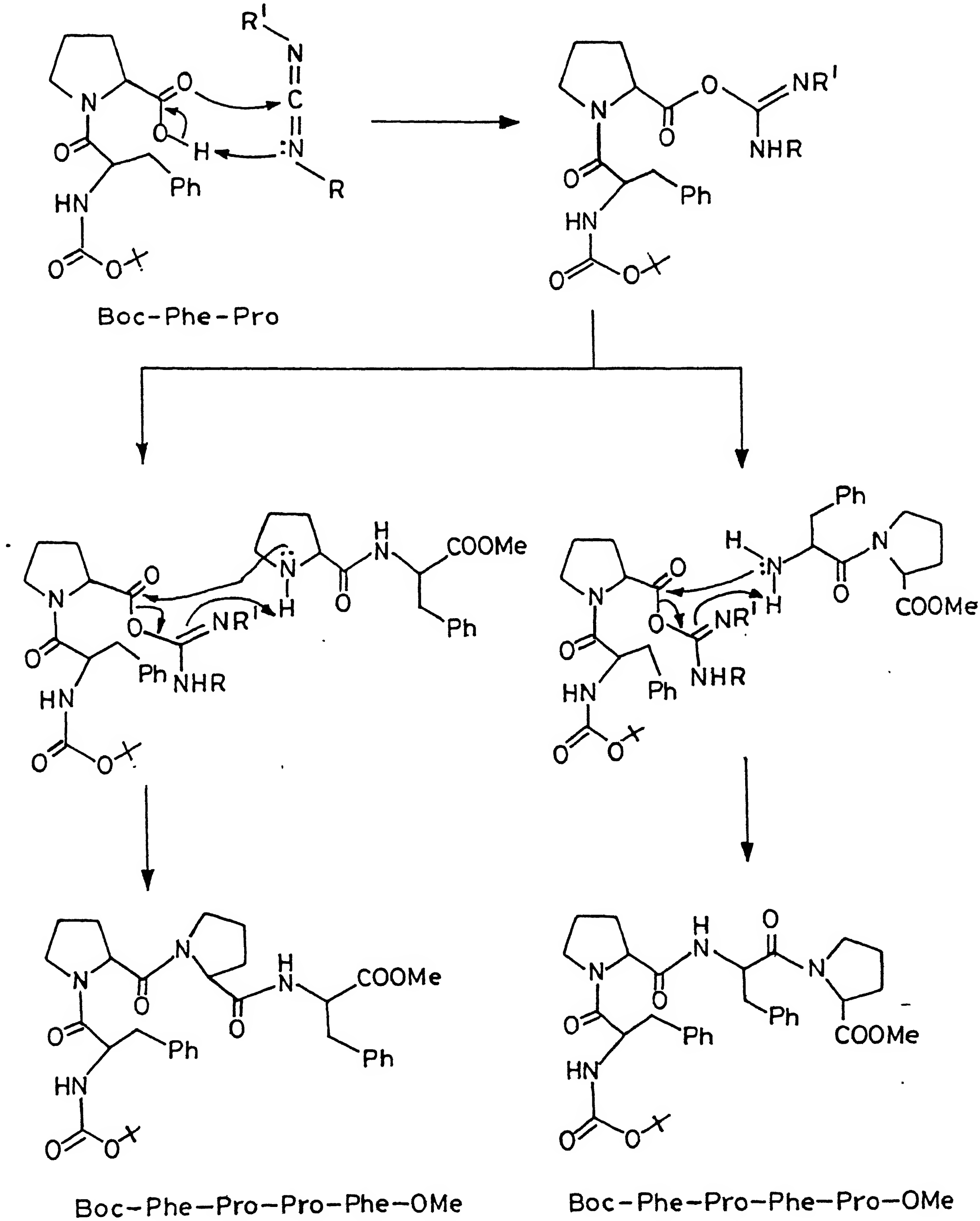
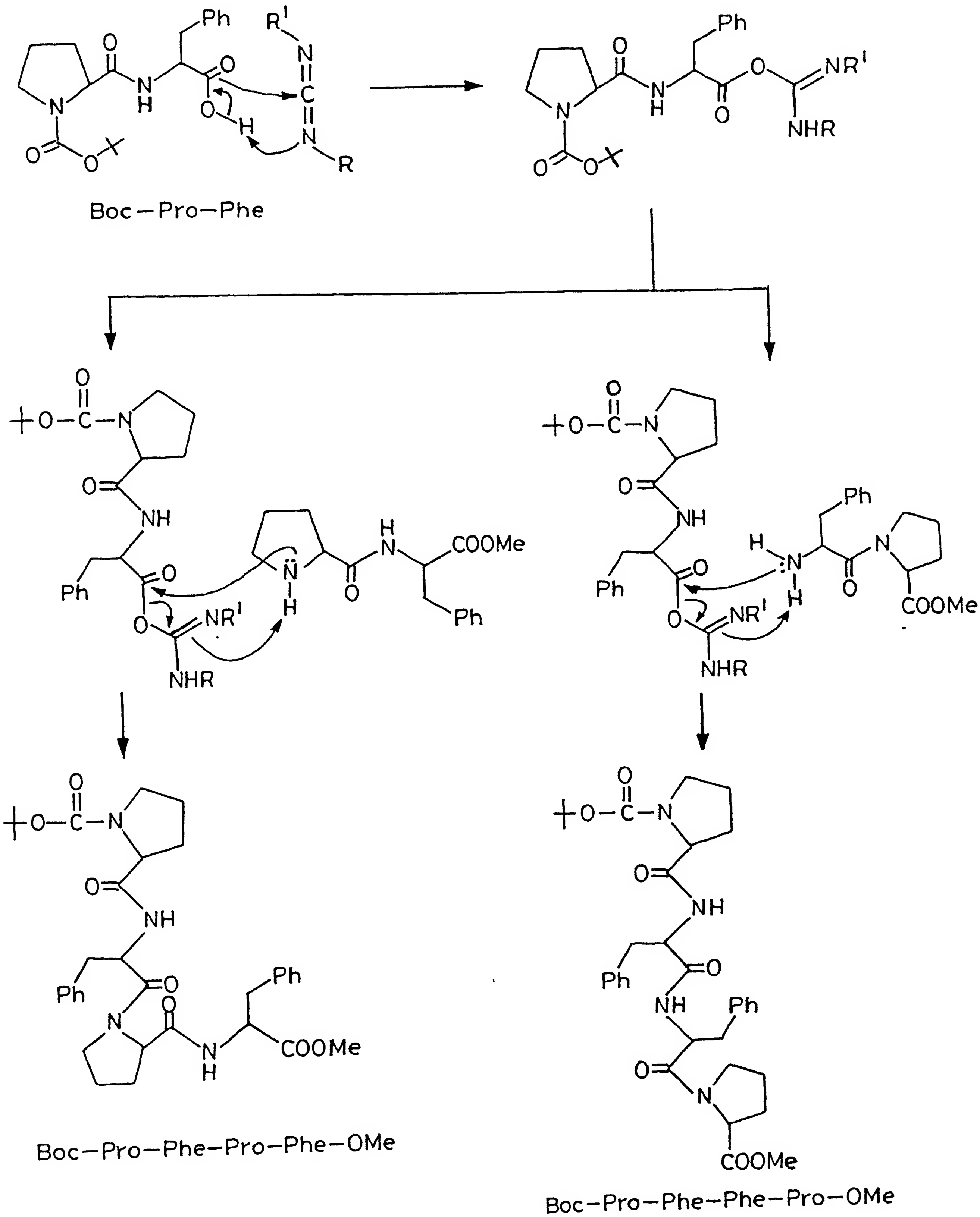


CHART.C.I.21

Major controls in Boc-Pro-Phe-Phe-Pro-OMe / Boc-Pro-Phe-Pro-Phe-OMe peptide bond formation mediated by water soluble carbodiimide in water.



decrease the propensity for peptide bond formation arising from an activated C terminal Pro residue. A complementary picture is presented in CHART.C.I.21 wherein possible pathways involved in the activation and subsequent peptidation of a C terminal Phe is shown.

A comparison of CHART.C.I.16 and CHART.C.I.21 would show that the pathways involved leading to amidation of the Phe terminal residue are quite similar.

It would be of interest to see whether some predictions can be made with respect to selectivity in tetrapeptide formation when equivalent amounts of Boc-Pro-Phe, Boc-Phe-Pro, Pro-Phe-OMe and Phe-Pro-OMe were allowed to react in water in presence of WSCDI, based on experimental findings previously reported here. At the outset, in view of the uniform difficulties seen in the peptide bond formation, it could be concluded that the presence of Boc-Phe-Pro-Pro-Phe-OMe in the mixture would be negligible. Again, based on the findings from the study using N/C blocked amino acids, it could be predicted that the activated Phe from Boc-Pro-Phe is likely to lead preponderantly to afford Boc-Pro-Phe-Phe-Pro-OMe in relation to Boc-Pro-Phe-Pro-Phe-OMe. Thus the only unpredictable factor here is the extent of formation of Boc-Phe-Pro-Phe-Pro-OMe which would be a function of steric difficulties with respect to Pro-Phe bond formation.

As the first step authentic samples of the four possible tetrapeptides that could arise from pathways shown in CHART.C.I.20 and CHART.C.I.21 were prepared and their HPLC profile with respect to clear cut differentiation was developed. TABLE.C.I.5 presents a profile of authentic samples.

In the event equivalent amounts of Boc-Phe-Pro, Boc-Pro-Phe, Phe-Pro-OMe and Pro-Phe-OMe were allowed to condense in presence of 2 equivalents WSCDI in water at pH 7.1. The reaction mixture was left stirred for 48 hours at room temperature, extracted with ethylacetate, dried, evaporated and the mixture was analyzed and compared with authentic tetrapeptide samples in HPLC.

In FIGURE.C.I.1e is presented the HPLC profile of the authentic tetrapeptides and

TABLE.C.I.5

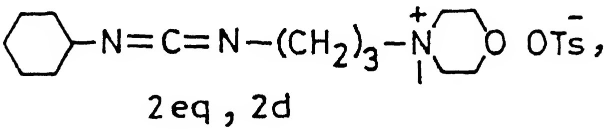
HPLC PROFILE OF AUTHENTIC N,C-PROTECTED TETRAPEPTIDES

Tetrapeptide	Mobile Phase	Flow Rate	Retention Time (Min)
Boc-Phe-Pro-Pro-Phe-OMe			10.88
Boc-Phe-Pro-Phe-Pro-OMe	MeCN : H_2O	0.8 mL/min	11.40
Boc-Pro-Phe-Pro-Phe-OMe	::		13.68
Boc-Pro-Phe-Phe-Pro-OMe	60 : 40		11.00

CHART.C.I.22

Experimentally determined preference for tetrapeptide formation in the reaction of Boc-Pro-Phe, Boc-Phe-Pro with Pro-Phe-OMe and Phe-Pro-OMe.

Boc-Phe-Pro	+	Phe-Pro-OMe	
1eq		1eq	$\xrightarrow{\text{(i)}}$
Boc-Pro-Phe	+	Pro-Phe-OMe	
1eq		1eq	

(i) 

Tetrapeptide	% Distribution	% Yield
Boc-Pro-Phe-Phe-Pro-OMe	77	58
Boc-Pro-Phe-Pro-Phe-OMe	23	17
Boc-Phe-Pro-Pro-Phe-OMe	0	-
Boc-Phe-Pro-Phe-Pro-OMe	0	-

FIGURE.C.I.1e

HPLC profile of authentic tetrapeptides for Proline and Phenylalanine

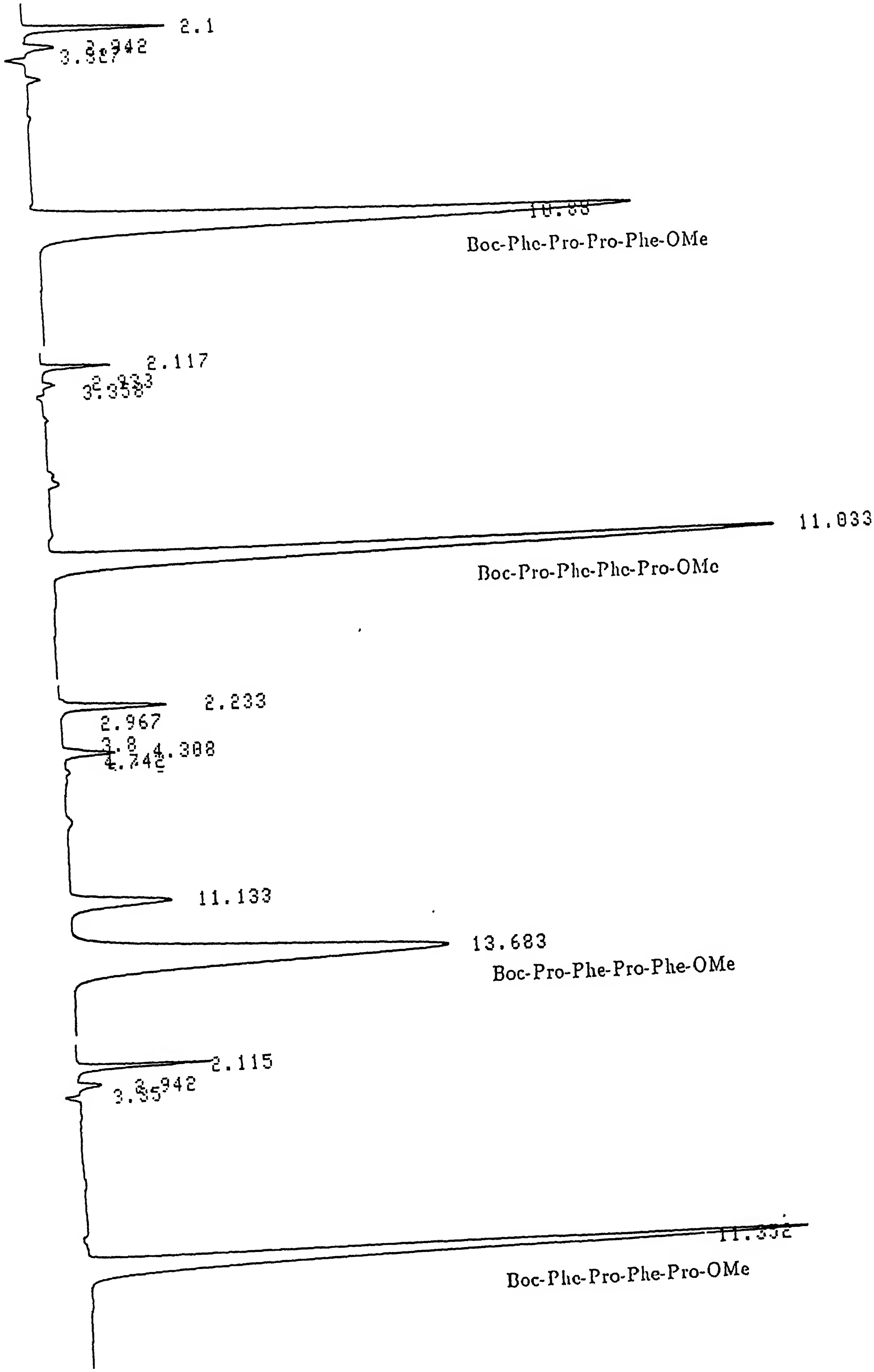
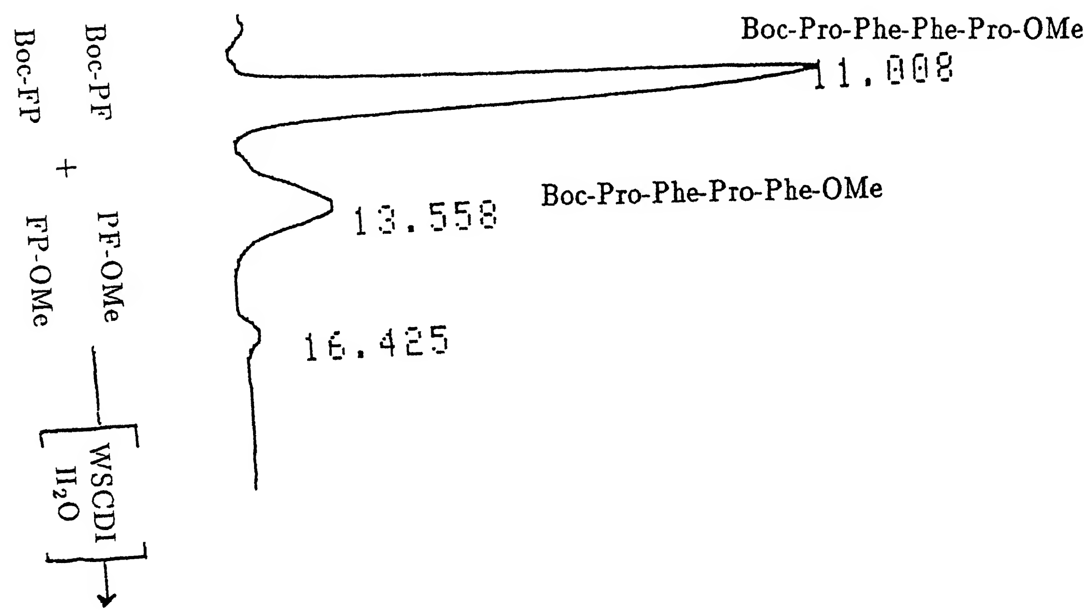


FIGURE.C.I.2e

Dipeptide distribution in the condensation of Boc-Pro-Phe and Boc-Phe-Pro with Pro-Phe-OMe and Phe-Pro-OMe in presence of WSCDI in water



in FIGURE.C.I.2e the analysis of the reaction mixture.

The experimentally determined preference for tetrapeptide formation is summarized in CHART.C.I.22. It could be seen from here that the preponderant product was the expected Boc-Pro-Phe-Phe-Pro-OMe (77%). The modest presence of Boc-Phe-Pro-Phe-Pro-OMe (23%) could have been anticipated from earlier studies, as expected Boc-Phe-Pro-Pro-Phe-OMe was absent. An unexpected feature was the absence of Boc-Pro-Phe-Pro-Phe-OMe and this could be readily rationalized on additional steric constraints (*vide supra*).

This experiment which demonstrated strong preferences in a blockwise mode condensation leading to tetrapeptide is of significance. Regardless of finer aspects of reaction mechanisms, selectivity would be expected in such reactions. The result here also provides encouragement to make peptides largely based on neighbour selectivity. This also supports a potential practical approach to peptide synthesis wherein sequence selectivity is achieved through repetitive cyclic operations. This can be illustrated with the example of Phe and Pro.

In the cycle of condensation involving Pro and Phe, Pro-Phe would be predominant, in the second cycle involving the unprotected dipeptide in reaction mixture Pro-Phe-Phe-Pro can be expected to be major product.

A wide domain has been traversed in this section. The theoretical and experimental probes employed here have confirmed the notion of selectivity in neighbour preferences, the ramification of which would have implications not only with respect to protein evolution and a better understanding of the profile of coded amino acid side chains but also towards the development of novel methodologies for peptide synthesis.

C.II. SEQUENCE ANALYSIS OF THE DNA RECOGNITION ELEMENT OF ZINC FINGER PROTEINS : A CASE FOR THE MODULAR APPROACH TO EVOLUTION OF INFORMATION-FUNCTION COMPOSITES

Detailed examination involving data base analysis and experimentation have shown that neighbour preferences in peptides are correlated to their intrinsic properties. This rationale makes it possible, the presence of sequence specific blocks of peptides. As stated previously, Darwinian Principle as natural selection when taken to its basic roots would mark the establishment of an interdependent relationship between informational system- nucleic acids - and functional system- proteins - as the starting point towards the protocell. According to this view if a particular peptide aided, however likely, in the synthesis of a oligonucleotide with a specific non random sequence, then the cycle of interdependence would have begun. The accumulation of the specific peptide would have led to the accumulation of more of the specific oligonucleotide. This aspect therefore forms a link between preferences amongst coded complement in the peptide bond formation and that pertaining to their interaction with the information system.

It is well possible to imagine that in the early stages of development leading to the protocell there could have existed numerous combinations of small composites consisting of blocks of nucleotides and coded amino acids. The assembly of these composites would then lead to the total information system capable of regenerating itself and via transcription and translation to protein ensemble to which it was originally linked. Thus it is logical to envisage that as in the case of peptides the relationship between nucleotides and amino acids would also exhibit, in this scenario, a non random behaviour. Interestingly, the genetic code (CHART.C.I.1) can be analyzed to delineate the possibility of existence of blocks of information function composites in the early stages of evolution leading to protocell. Across the living kingdom the genetic code enjoys an extraordinary degree of fidelity in terms of direction of functional and structural proteins in an

error free manner and thus, apart from minor aberrations, the code represents a true relationship between the functional and informational system. This is logical since even minor variations from the code could lead to functional systems of altered sequences and therefore altered structures which would be catastrophic to life. Granted that once the code is chosen, the option for aberration is highly restricted, it should be possible to obtain a direct correlation between the degeneracy of the code to the distribution of coded amino acid residues in proteins. In this event, the functional proteins that are present today arose subsequent to the freezing of the genetic code. A linear relationship between the percentage of each of 20 coded amino acids expected on the basis of the degeneracy of the code through which they are actually present in the data base must be observed. On the other hand is a situation where the functional system initially arose from an optimal composite with the corresponding functional system by a blockwise addition and aggrandization, such a correlation need not be adhered to. Thus a non random profile in the distribution of coded amino acids compared to that expected on the basis of degeneracy of the code would support the notion that methodologies for error free replication and translation were invoked prior to freezing of the code. In terms of present day observations the readings of a triplet sequence in its structural element comprises of much larger than one amino acid side chain as exemplified with zinc fingers. The implications are clear that information function composites which recognizes even one amino acid residue must have a unit auxiliary residue to promote such elements. The primary data base consisting of 9416 residues (TABLE.C.I.1) was analyzed from this vantage. The entire set was analyzed in terms of occurrence of each of the 20 coded amino acids. The observed percentage were computed from this number. The expected percentage on the basis of the degeneracy i.e observed in the genetic code was computed using the following equation :

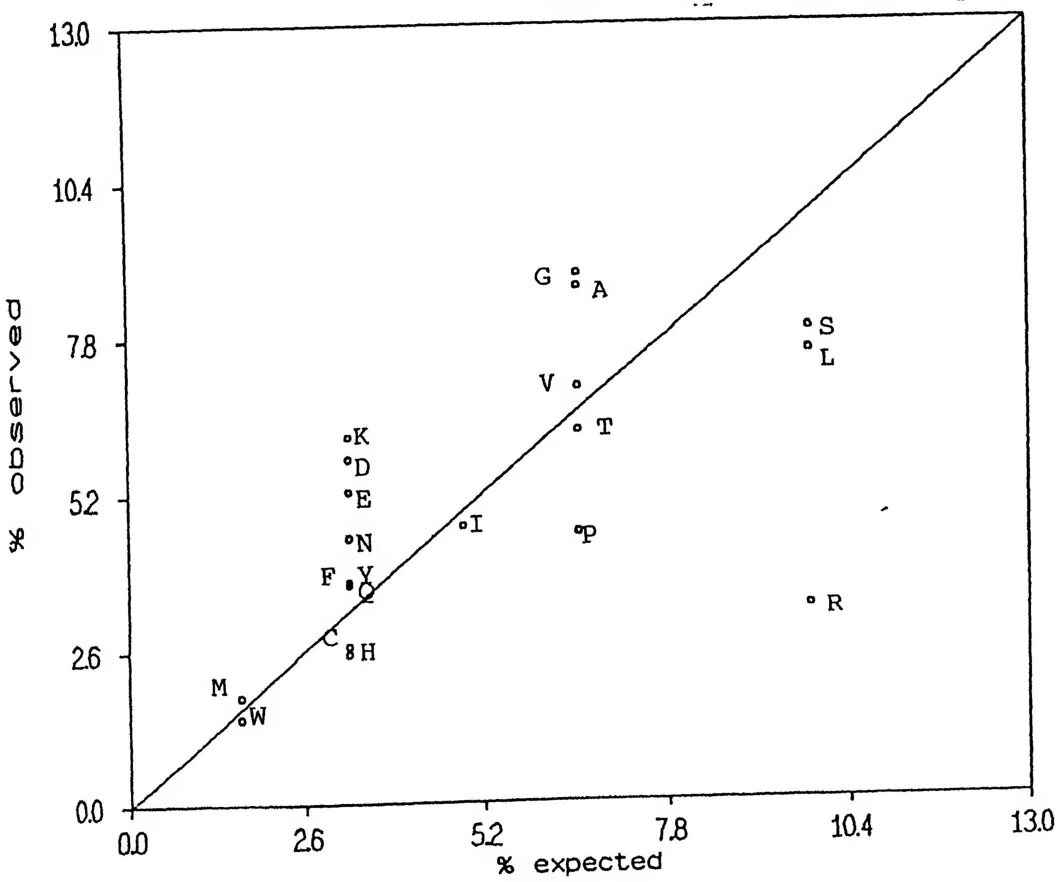
Expected percentage = (n x 100) / 61, where n is the number of degeneracy of the particular amino acid, in the genetic code.

In CHART.CII.1, the percentage expected for each of the residues is plotted against that actually observed. As stated previously, in the event the observed values match with expected one, all the 20 residues would lie along the diagonal element. An examination of CHART.CII.1 would show that most of the coded amino acids are far from ideal, thus strongly implicating that the information system that evolved, had the ability to produce copies by error free replication and to construct the functional system via transcription process before the code was frozen. Yet another interesting aspect is that protein-DNA recognition can be readily accomplished by secondary structural elements principally α-helix structure. This is true both in terms of prokaryotic and eukaryotic systems. The involvement of β- sheet in DNA recognition is much less present. In CHART.C.II.2 and CHART.C.II.3 are shown, respectively, the percentage expected occurrence against that observed, pertaining to α-helix and β-sheet, and, as expected, the variations can easily be related to the structural needs associated with the construction of such secondary structural elements. CHART.C.II.4 endeavours to explain graphically what has been stated above. The information function composites could arise either by sequential addition of units (digital approach) or by the joining of modular blocks (modular approach), the latter being most efficient. Once a satisfactory information system is created, the functional counterpart becomes redundant and informational system alone, which can be represented as gene or a segment of DNA, could direct replication and transcription by the digital mode, which now becomes more efficient. CHART.C.II.5 summarizes these events and it could be seen from here that once the master template gene is constructed and so recognized, the only modular interaction that is seen presently, pertains to modular recognition of DNA sequence which is indeed a prerequisite for initiation of transcription.

With this background it was considered to be of importance to perform a sequence analysis that are present in DNA recognition elements of zinc-finger proteins to assess the correlation pertaining to neighbour with those which are recognized in the present

CHART.C.II.1

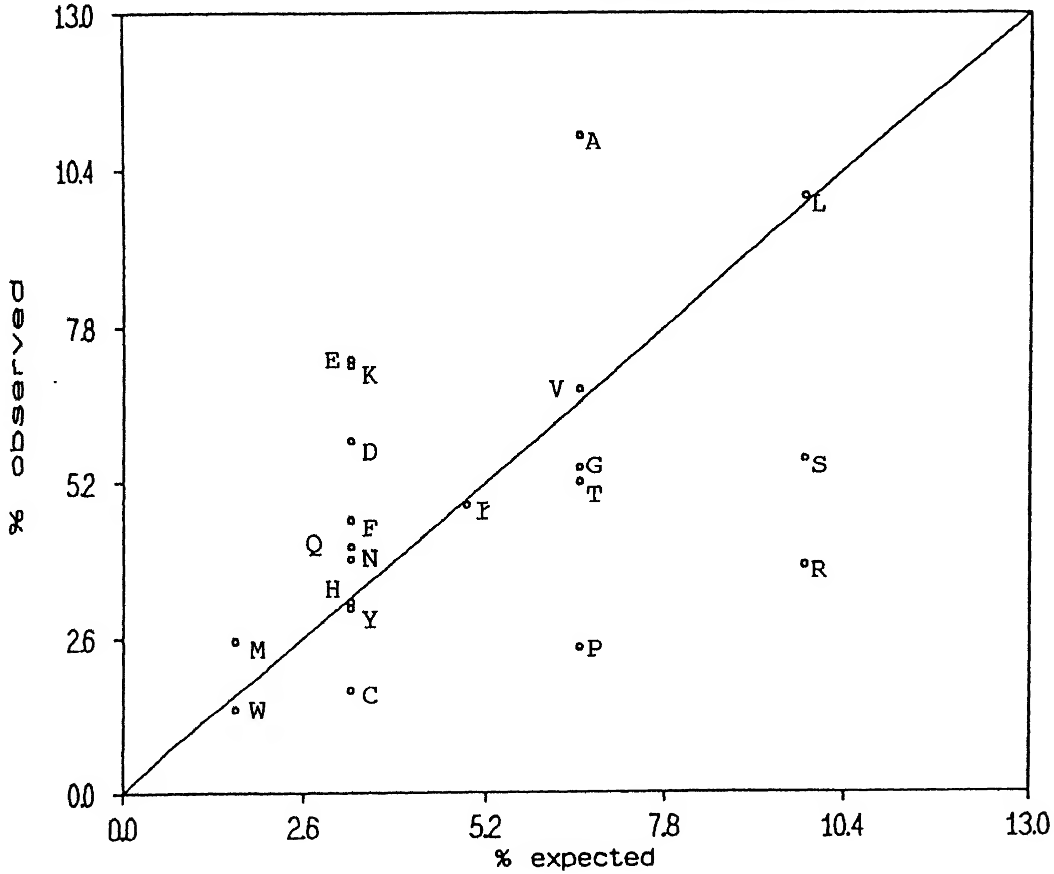
Distribution of amino acids found in 62 proteins vs. their expected frequencies if codons occurred randomly



Amino Acid Residue	Total No.	%Observed	%Expected
Alanine	815	8.65	6.55
Cysteine	252	2.67	3.27
Aspartic Acid	545	5.78	3.27
Glutamic Acid	495	5.25	3.27
Phenylalanine	354	3.75	3.27
Glycine	835	8.86	6.55
Histidine	239	2.53	3.27
Isoleucine	439	4.66	4.91
Lysine	579	6.14	3.27
Leucine	709	7.52	9.83
Methionine	172	1.82	1.63
Asparagine	421	4.47	3.27
Proline	427	4.53	6.55
Glutamine	349	3.70	3.27
Arginine	309	3.28	9.83
Serine	745	7.91	9.83
Thereonine	586	6.22	6.55
Valine	656	6.96	6.55
Tryptophan	139	1.47	1.63
Tyrosine	350	3.71	3.27
Total	9416		

CHART.C.II.2

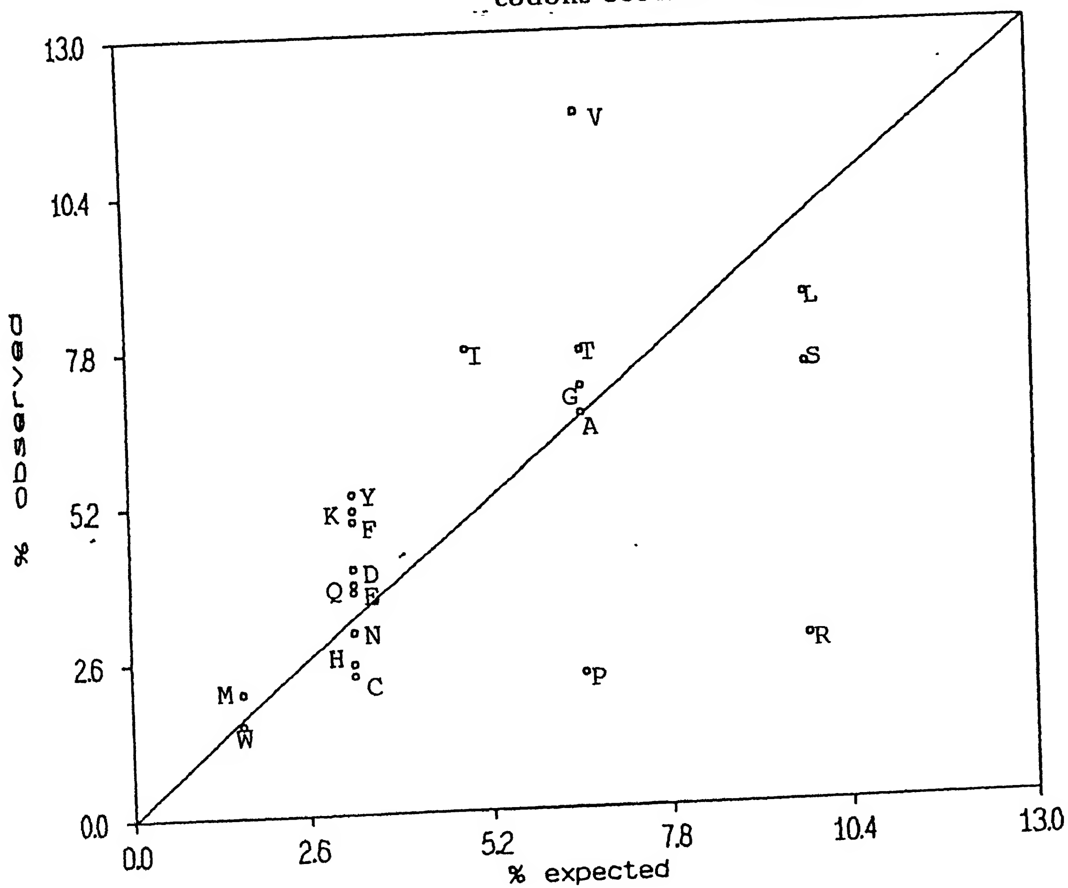
Distribution of amino acids found in α-helix vs. their expected frequencies if codons occurred randomly



Amino Acid Residue	Total No.	%Observed	%Expected
Alanine	330	10.96	6.55
Cysteine	052	1.72	3.27
Aspartic Acid	178	5.91	3.27
Glutamic Acid	219	7.27	3.27
Phenylalanine	138	4.58	3.27
Glycine	164	5.45	6.55
Histidine	094	3.19	3.27
Isoleucine	145	4.81	4.91
Lysine	216	7.17	3.27
Leucine	300	9.97	9.83
Methionine	077	2.55	1.63
Asparagine	118	3.92	3.27
Proline	074	2.45	6.55
Glutamine	124	4.12	3.27
Arginine	115	3.82	9.83
Serine	169	5.61	9.83
Thereonine	157	5.21	6.55
Valine	204	6.77	6.55
Tryptophan	042	1.39	1.63
Tyrosine	093	3.09	3.27
Total	3009		

CHART.C.II.3

Distribution of amino acids found in in β-sheet vs. their expected frequencies if codons occurred randomly



Amino Acid Residue	Total No.	%Observed	%Expected
Alanine	100	6.60	6.55
Cysteine	038	2.51	3.27
Aspartic Acid	062	4.09	3.27
Glutamic Acid	056	3.70	3.27
Phenylalanine	074	4.89	3.27
Glycine	107	7.07	6.55
Histidine	035	2.31	3.27
Isoleucine	117	7.73	4.91
Lysine	077	5.08	3.27
Leucine	129	8.52	9.83
Methionine	031	2.04	1.63
Asparagine	046	3.04	3.27
Proline	034	2.24	6.55
Glutamine	058	3.83	3.27
Arginine	042	2.77	9.83
Serine	111	7.33	9.83
Thereonine	116	7.66	6.55
Valine	176	11.63	6.55
Tryptophan	023	1.52	1.63
Tyrosine	081	5.35	3.27
Total	1513		

CHART.C.II.4

Options for Information-Function Composites

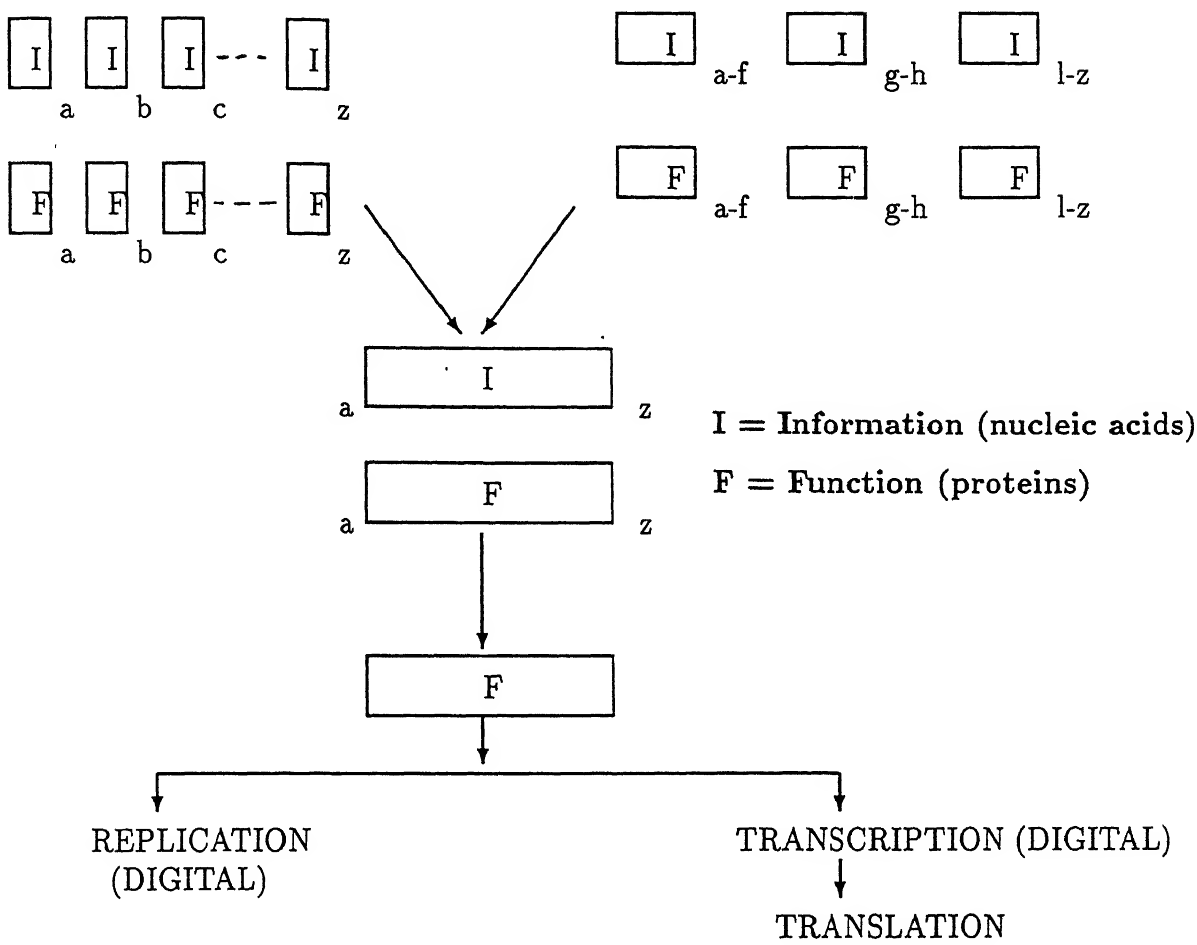
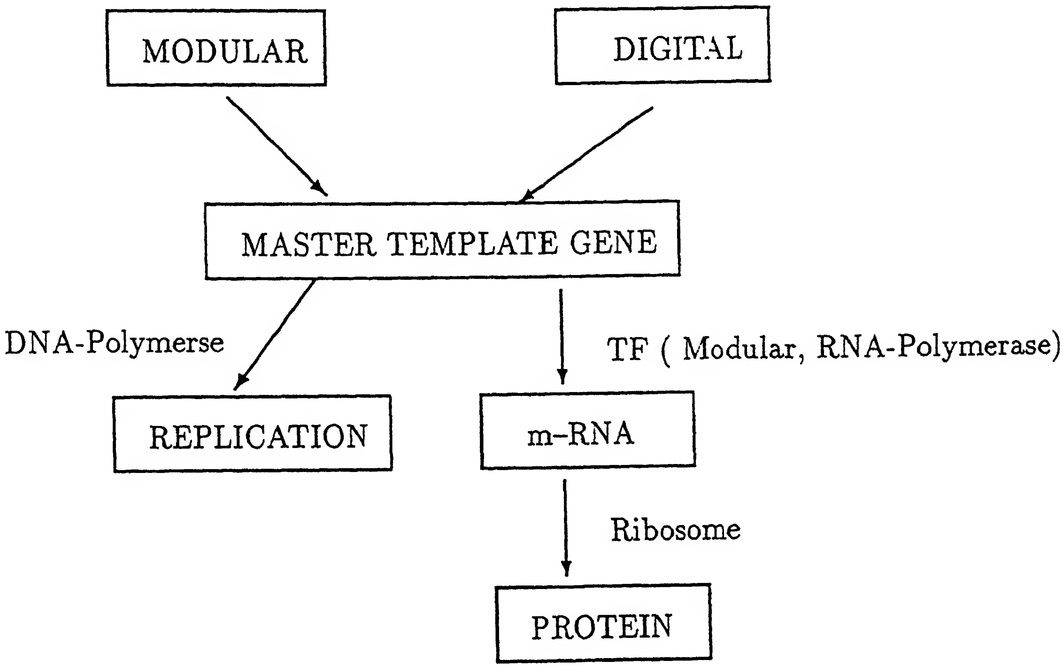


CHART.C.II.5

Stabilization of information-function relationship



day functional systems. The logic here is that in the event present day functional systems harbour imprints of early evolution and that this process is associated with a modular approach, then a reasonable correlation can be expected between a consensus zinc finger recognition element actually present with the one constructed from the data base generated from functional proteins.

The study of DNA-Protein interactions is currently the focus of multifaceted endeavours, not only because of the potential utility in diverse domains, but also because this phenomena perhaps symbolizes the perfection pertaining to molecular recognition. Indeed, the phenomenon of DNA-Protein recognition could be identified as a major step pertaining to cellular organization that demands error free replication of the information system and the subsequent translation to generate precisely defined functional and structural proteins. Whilst the importance of DNA-Protein recognition pertaining to early stages of evolution has been recognized, the mechanisms by which this could have manifested itself remains unclear. The major difficulty here is that, as is true with many aspects of evolution, the pathways by which the present day manifestations came into being are largely obliterated. The DNA-Protein interaction is associated with a number of cellular processes. The ribosomal protein synthesis is often considered as an example where sequences in the information system are matched with sequences of the functional protein as directed by genetic code. In the ribosomal protein synthesis the information system, represented by m-RNA, recognizes the amino acid component of the functional system in an individual mode and thus the protein is constructed by addition of individual units, based on single codon-recognition. The unit recognition and attachment mode here for the construction of proteins is an inferior strategy wherein modular units that form part of the composite could be joined to generate the desired functional system. The dichotomy here can be resolved with the notion that with the construction of the information templates that can direct error free transcription and translation operations, it would naturally be more practical, in terms of ready availability of complementary

slots, singular rather than modular blocks and thus replication as well as transcription adopts this mode, since the option for either a unit or a modular pathway is available only till the construction of the information template (CHART.C.II.4, CHART.CII.5). Interestingly, when an information template is absent, as in the case with non-ribosomal peptide synthesis, a group of enzymes forms a composite wherein each unit is able to put together a constituent module (SECTION.C.III).

Leads that suggest a block wise (modular) approach in the genesis of the early information-function composite are worthwhile from the vantage of (1) a deeper understanding of the genetic code (2) role of organic chemistry of coded amino acid side chains (3) the design of consensus sequences that reflects a particular pattern of function (4) the delineation of the option pertaining to replace the amino acid by cassette mutagenesis (5) protein evolution and structure and (6) the construction of molecules based on Protein-DNA recognition.

A correlation of the sequence profile of proteins that are known to recognize DNA on a modular basis with that from a broad data base generated from representative protein sequences should provide clues pertaining to early function-information recognition. Thus, the zinc-finger proteins, originally recognized as transcription factors and which, by all the evidences available including X-ray crystallography appear to recognize the DNA sequences on a modular basis, presented themselves as the best available models to probe the genesis of information-function composite on a modular basis or otherwise.

Zinc fingers conform to an approximately 30 residue peptide sequence that can be represented as ...$F(Y)XCX_{2-4}CX_3FX_5LX_2HX_3HX_5$... (F = Phenylalanine, Y=Tyrosine, C=Cysteine, L=Leucine, H=Histidine, X=number of variable residues) and generally comprising of two β- strand like structures followed by an α-helix anchored on Zn^{II} template coordinated to the pairs of cysteine and histidine. Modular arrays of zinc fingers bind to the major groove of B-DNA and wind around it such that the movement of one finger to the next involves a rotation of *ca* 96^0 around the DNA axis, and a

translation of *ca* 10 A^0 along the DNA axis, the projection corresponding to a 3 base pair contact usually involving the amino acid residue that immediately precedes the α-helix as well as the second, third and sixth residues of the α- helix. Most contacts are with the guanine rich strand of DNA[20-22].

Till today well over 200 zinc finger genes have been reported and it has been estimated that there are several hundred zinc finger encoding genes in the human genome[23,24]. The large number of zinc finger sequences provides an opportunity to learn more about their structure and functions through their sequences. A most recent example of this, based on a detailed sequence analysis is the possibility for construction of *de novo* zinc fingers with predictable DNA base sequence contact[25]. A generalized profile of zinc finger module is presented in CHART.C.II.6.

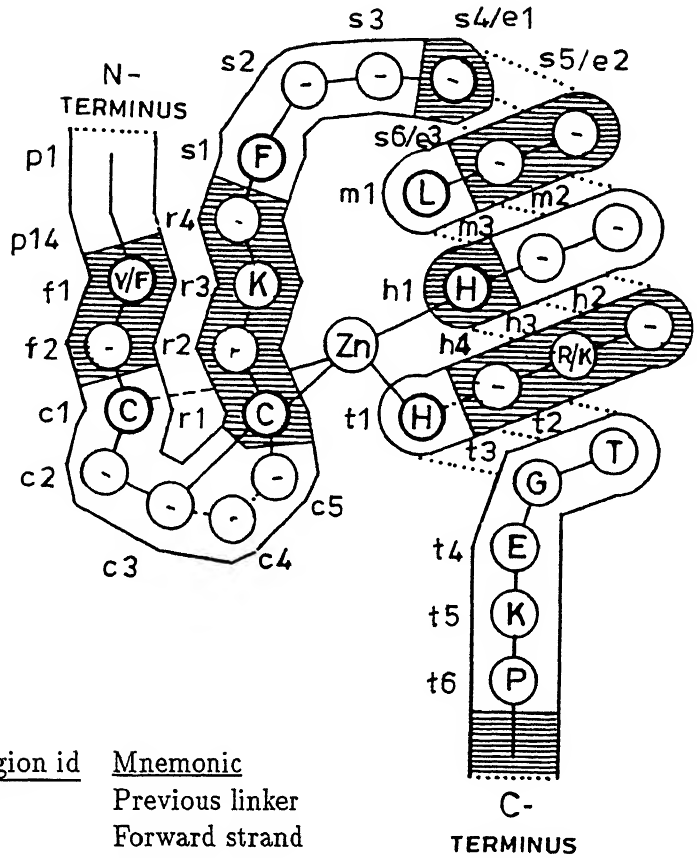
As could be seen from CHART.C.II.6, each module of the zinc finger comprises of 2 anti parallel β- sheet like structures followed by an α-helix. This structural ensemble is anchored on the Zn template arising from tetrahedral coordination of pairs of conserved cysteine and histidine. The recognition loops that span the middle cysteine and histidine, the focus of the present analysis, have a typical profile as shown in CHART.C.II.7. Thus, starting from a conserved phenylalanine residue, the first contact amino acid, placed just precedes the α-helix, is reached via a single variable residue. The second contact residue, well within the α-helix is linked to the first one by two variable residues. A conserved leucine follows the second contact amino acid and leads to the third contact residue via a single variable unit. The recognition loop is completed by the direct linking of a third contact residue to the conserved histidine.

The objective of the present analysis was then to generate a consensus sequence of a typical zinc finger recognition loop from á preference data base arrived from functional protein and the match with what is determined.

In TABLE.C.II.1 is presented a listing of 135 zinc finger modules used in the construction of the data base. This table also shows the number of zinc fingers present in

CHART.C.II.6

A GENERAL PROFILE OF A ZINC FINGER



Region id	Mnemonic
p	Previous linker
f	Forward strand
c	Cysteine loop
r	Reverse strand
s	Shoulder
m	Main helix
e	Entry helix
h	Histidine helix
t	Tail (following linker)

A schematic diagram of a zinc finger is shown. The finger is divided into nine regions, eight of which begin with a conserved amino acid position (the exception is the *e* region). Conserved positions are circled in bold. Each region is assigned a single letter region identifier (*p,f,...t*). A position is referred to by its region identifier followed by the position within the region. Position 1 of a region is the conserved position the region starts with. Each region identifier has a mnemonic value, as shown. Note that the *e* region overlaps with the *s* region. The *e1* position is the same as the *s4* position.

CHART.C.II.7

DNA CONTACT ELEMENT OF ZINC FINGER MODULE

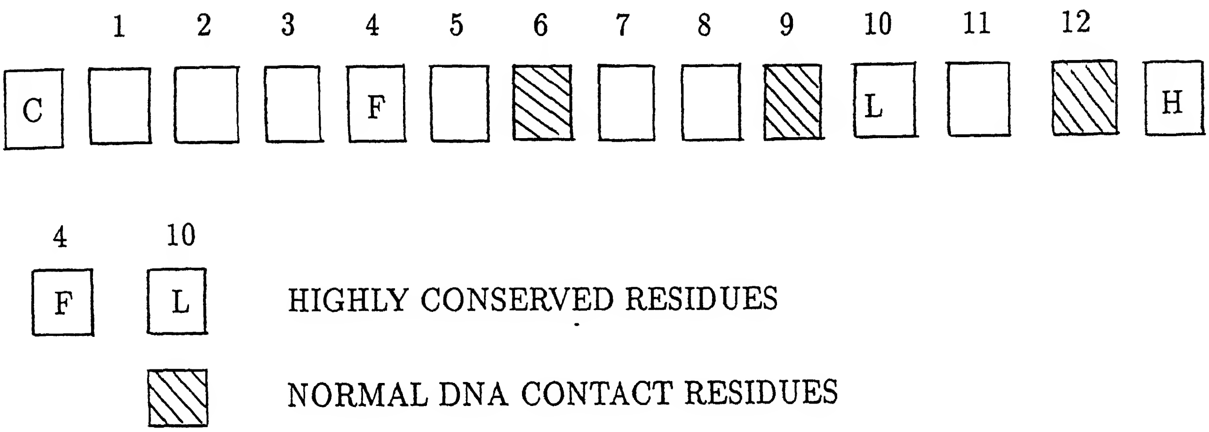


TABLE.C.II.1

Listing of zinc finger proteins used in the construction of the data base

No.	Name	No of Fingers Present
1	Human Spl	3
2	Drosophila Serendipity β	6
3	Drosophila Serendipity δ	7
4	Drosophila Kruppel	5
5	Drosophila Snail	5
6	Xenopus Xfin	36
7	Drosophila Terminus	1
8	Yeast SW15	3
9	Xenopus Transcription Factor IIIA	9
10	Yeast ADRI	2
11	Drosophila Krh	6
12	Mouse Mkrl	7
13	Mouse Mkr2	9
14	Drosophila Hunchback	6
15	Trypanosome TRS-1	5
16	Mouse NGFI-A	3
17	Human ZFY	13
18	Xenopus p43 5S RNA Binding Protein	9

each of the proteins analyzed. Thus the total number of zinc fingers analyzed in the present study is 135. The sequences of the recognition module which forms the focus of the present analysis are listed as APPENDIX.C.II.1.

Using the two programs, used in the previous section C.I, namely, that for delineation of neighbour preferences and the subsequent one for deriving the "Left-Right" preferences pertaining to a central residue, in tandem, a profile presented in CHART.C.II.8 was derived for the entire set of zinc finger modules. Pertinent to the present study is the nature of such preferences which are actually present in recognition modules in zinc fingers. Of the 3568 residues which are present in 135 zinc finger modules 1890 (52 %) form a part of its DNA recognition element. These were also subjected to a tandem analysis and the results are shown in CHART.C.II.9.

The formation of information function composite involving individual zinc finger modules and the corresponding nucleotide sequences should be termed as a very significant event in evolution towards a protocell. Here, again it is possible to secure some idea about the formation of such modular composites within the time frame wherein the markers is the freezing of the genetic code. Thus charts similar to CHART.C.II.1, CHART.C.II.2, CHART.C.II.3 were constructed using data base pertaining to the total module (CHART.C.II.10) and recognition region CHART.C.II.11.

A gross comparison of CHART.C.II.1, CHART.C.II.2, CHART.C.II.3 with CHART. C.II.10 and CHART.C.II.11 show that in the construction of the zinc finger modules the selectivity criterion is most stringent as evident by a much wider scattering around the ideal diagonal element reference. Even more interesting would be a comparison between CHART.C.II.10 and CHART.C.II.11, particularly in the light of various unique structural features that have been highlighted during the last decade pertaining to the profile of the approximately 30 residues of zinc finger modules not only in terms of secondary structure but also, more importantly in terms of the mechanism associated with direct contact recognition between the amino acid side chains and the code bases.

CHART.C.II.8

MASTER ANALYSIS FOR NEIGHBOUR (LEFT-RIGHT) PREFERENCE

(Zinc Finger Modules, 3568 Residues)

NEIGHBOUR	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	04/04	08/07	07/06	05/02	14/13	07/03	09/09	02/01	06/17	17/06	02/02	04/01	01/02	02/02	06/05	06/14	04/05	06/07	00/01	02/05
C	07/08	03/03	43/12	23/45	05/02	50/17	12/15	00/21	22/37	07/10	03/04	19/03	30/15	13/29	10/17	28/10	18/19	05/21	00/00	01/11
D	06/07	12/43	02/02	08/01	03/00	04/00	07/05	04/02	19/05	23/02	00/00	01/01	03/03	01/04	07/07	06/25	02/02	03/03	00/02	03/00
E	02/05	45/23	01/08	05/05	03/07	08/06	07/11	02/03	14/09	09/09	03/02	02/06	07/05	02/06	14/03	05/11	03/11	07/04	00/00	08/13
F	13/14	02/05	00/03	07/03	03/03	02/16	07/06	17/01	21/16	04/03	03/01	06/03	04/25	07/06	11/14	29/33	22/18	16/06	01/00	01/00
G	03/07	17/50	00/04	06/08	16/02	00/00	03/04	03/01	38/14	09/00	03/02	01/01	50/05	02/03	06/02	02/08	02/53	06/02	00/00	04/05
H	09/09	15/12	05/07	11/07	06/07	04/03	14/14	19/32	33/58	35/21	20/06	06/11	04/08	42/06	17/39	30/21	65/55	11/30	05/01	04/08
I	01/02	21/00	02/04	03/02	01/17	01/03	32/19	03/03	09/08	04/18	01/03	01/03	02/04	17/01	12/21	04/02	01/05	04/04	00/00	01/01
K	17/06	37/22	05/19	09/14	16/21	14/38	58/33	08/09	28/28	15/29	07/09	09/09	09/05	05/25	28/15	35/15	26/17	14/19	00/04	09/12
L	06/17	10/07	02/23	09/09	03/04	00/09	21/35	18/04	29/15	17/17	04/03	04/19	03/03	11/05	28/05	13/21	18/10	24/09	00/03	05/07
M	02/02	04/03	00/00	02/03	01/03	02/03	06/20	03/01	09/07	03/04	00/00	04/02	01/00	00/00	11/04	01/03	01/00	02/00	00/00	03/00
N	01/04	03/19	01/01	06/02	03/06	01/01	11/06	03/01	09/09	19/04	02/04	00/00	03/01	04/11	06/04	16/13	03/05	04/04	02/00	03/05
P	02/01	15/30	03/03	05/07	25/04	05/50	08/04	04/02	05/09	03/03	00/01	01/03	00/00	01/02	04/07	06/03	05/01	03/04	02/00	45/08
Q	02/02	29/13	04/01	06/02	06/07	03/02	06/42	01/17	25/05	05/11	00/00	11/04	02/01	05/05	28/08	10/12	08/06	03/06	00/00	02/11
R	05/06	17/10	07/07	03/14	14/11	02/06	39/17	21/12	15/28	05/28	04/11	04/06	07/04	08/28	19/19	38/20	33/13	05/04	02/02	08/10
S	14/06	10/28	25/06	11/05	33/29	08/02	21/30	02/04	15/35	21/13	03/01	13/16	03/06	12/10	20/38	27/27	09/04	15/03	03/02	05/05
T	05/04	19/18	02/02	11/03	18/22	53/02	55/65	05/01	17/26	10/18	00/01	05/03	01/05	06/08	13/33	04/09	08/08	06/02	03/02	04/14
V	07/06	21/05	03/03	04/07	06/16	02/06	30/11	04/04	19/14	09/24	00/02	04/04	04/03	06/03	04/05	03/15	02/06	06/06	00/00	07/01
W	01/00	00/00	02/00	00/00	00/01	00/00	01/05	00/00	04/00	03/00	00/00	00/02	00/02	00/00	02/02	02/03	02/03	00/00	00/00	01/00
Y	05/02	11/01	00/00	13/08	00/01	05/04	08/04	01/01	12/09	07/05	00/03	05/03	08/45	11/02	10/08	05/05	14/04	01/07	00/01	00/00

CENTRAL RESIDUE

Read out format: Central residue X can have neighbour profile YX(left preference) or XY(right preference) example for P, FP:PF :: 04:25

CHART.C.II.9

MASTER ANALYSIS FOR NEIGHBOUR (LEFT-RIGHT) PREFERENCE

(DNA contact region of zinc finger modules, 1890 Residues)

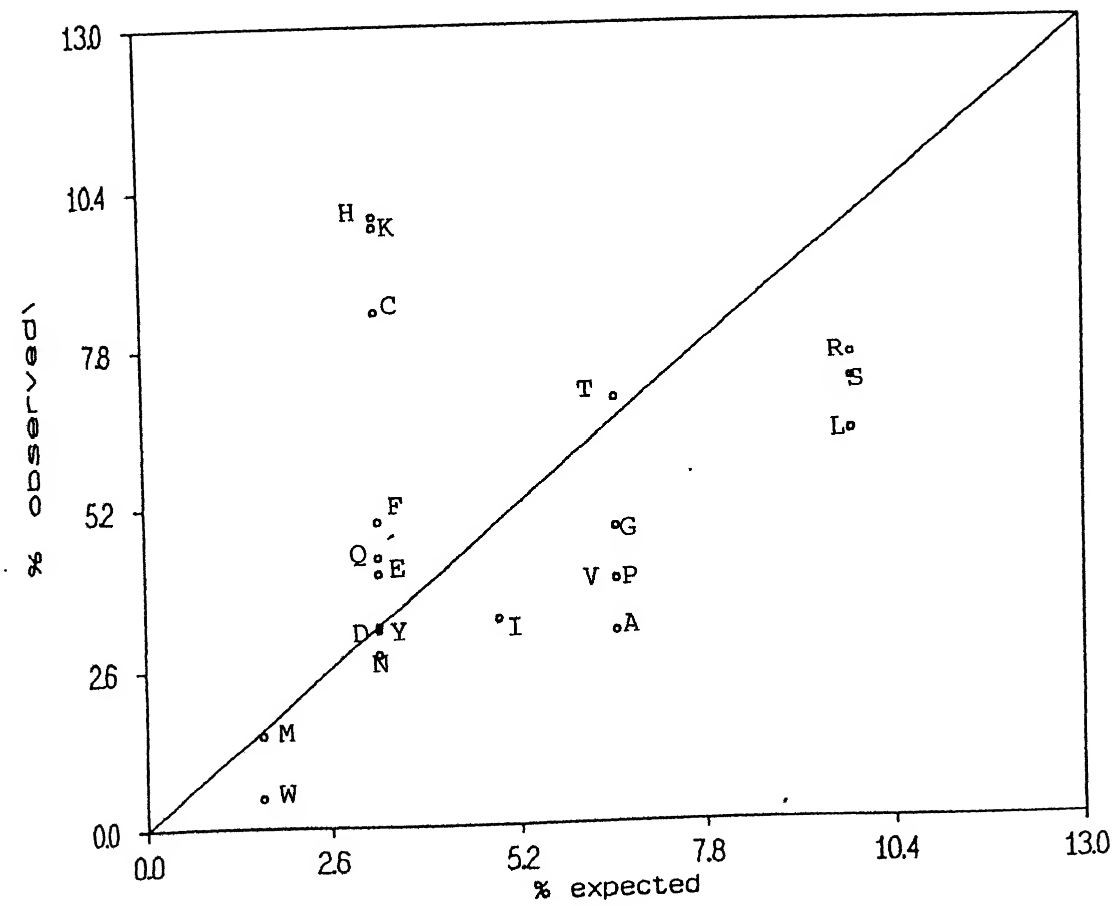
NEIGHBOUR	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	01/01	00/03	03/04	02/01	13/09	02/03	07/01	01/01	04/14	16/03	02/00	03/00	00/01	00/01	03/04	06/14	02/04	03/04	00/01	02/01
C	03/01	00/00	29/00	10/00	04/00	43/01	05/67	00/00	09/04	02/00	01/00	07/00	06/00	10/03	05/02	07/01	05/01	02/00	00/00	00/01
D	04/03	00/29	01/01	05/01	03/00	02/00	07/00	01/01	14/03	20/01	00/00	01/00	01/00	00/02	07/06	03/23	01/01	00/02	00/00	03/00
E	01/02	00/10	01/05	01/01	03/02	02/03	04/00	00/02	08/04	08/05	01/00	01/01	00/00	01/03	11/03	02/03	01/07	03/01	00/00	05/01
F	09/13	00/04	00/03	02/03	03/03	02/14	02/03	17/01	14/13	04/02	02/01	05/02	01/05	02/03	08/12	28/28	20/17	11/05	01/00	01/00
G	03/02	01/43	00/02	03/02	14/02	00/00	01/00	00/01	36/12	07/00	00/01	00/00	01/02	01/02	04/01	01/01	01/02	02/02	00/00	01/01
H	01/07	67/05	00/07	00/04	03/02	00/01	11/11	01/04	07/50	09/13	01/02	03/09	02/00	04/05	03/29	04/11	00/14	01/11	01/00	00/01
I	01/01	00/00	01/01	02/00	01/17	01/00	04/01	02/02	05/03	02/14	00/00	01/01	00/00	14/01	04/00	02/00	01/02	03/01	00/00	00/00
K	14/04	04/09	03/14	04/08	13/14	12/36	50/07	03/05	23/23	07/25	05/05	05/08	03/03	01/16	21/11	32/09	18/11	10/17	00/03	02/02
L	03/16	00/02	01/20	05/08	02/04	00/07	13/09	14/02	25/07	13/13	03/01	04/15	01/00	08/04	14/03	08/20	17/06	18/08	00/03	04/05
M	00/02	00/01	00/00	00/01	01/02	01/00	02/01	00/00	05/05	01/03	00/00	02/01	00/00	00/00	03/01	01/03	01/00	00/00	00/00	03/00
N	00/03	00/07	00/01	01/01	02/05	00/00	09/03	01/01	08/05	15/04	01/02	00/00	00/01	04/10	04/03	16/10	02/03	02/03	02/00	00/05
P	01/00	00/06	00/01	00/00	05/01	02/01	00/02	00/00	03/03	00/01	00/00	01/00	00/00	00/00	02/05	04/01	01/00	02/01	01/00	00/00
Q	01/00	03/10	02/00	03/01	03/02	02/01	05/04	01/14	16/01	04/08	00/00	10/04	00/00	03/03	09/06	08/09	03/04	01/05	00/00	00/02
R	04/03	02/05	06/07	03/11	12/08	01/04	29/03	00/04	11/21	03/14	01/03	03/04	05/02	06/09	11/11	35/15	03/10	00/02	01/00	02/02
S	14/06	01/07	23/03	03/02	28/28	01/01	11/04	00/02	09/32	20/08	03/01	10/16	01/04	09/08	15/35	23/23	07/04	07/03	03/01	02/02
T	04/02	01/05	01/01	07/01	17/20	02/01	14/00	02/01	11/18	06/17	00/01	03/02	00/01	04/03	10/03	04/07	07/07	04/02	02/02	02/07
V	04/03	00/02	02/00	01/03	05/11	02/02	11/01	01/03	17/10	08/18	00/00	03/02	01/02	05/01	02/00	03/07	02/04	03/03	00/00	02/00
W	01/00	00/00	00/00	00/00	00/01	00/00	00/01	00/00	03/00	03/00	00/00	00/02	00/01	00/00	00/01	01/03	02/02	00/00	00/00	01/00
Y	01/02	01/00	00/03	01/05	00/01	01/01	01/00	00/00	02/02	05/04	00/03	05/00	00/00	02/00	02/02	02/02	07/02	00/02	00/01	00/00

CENTRAL RESIDUE

Read out format: Central residue X can have neighbour profile YX(left preference) or XY(right preference) example for P, FP:PF :: 01:05

CHART.C.II.10

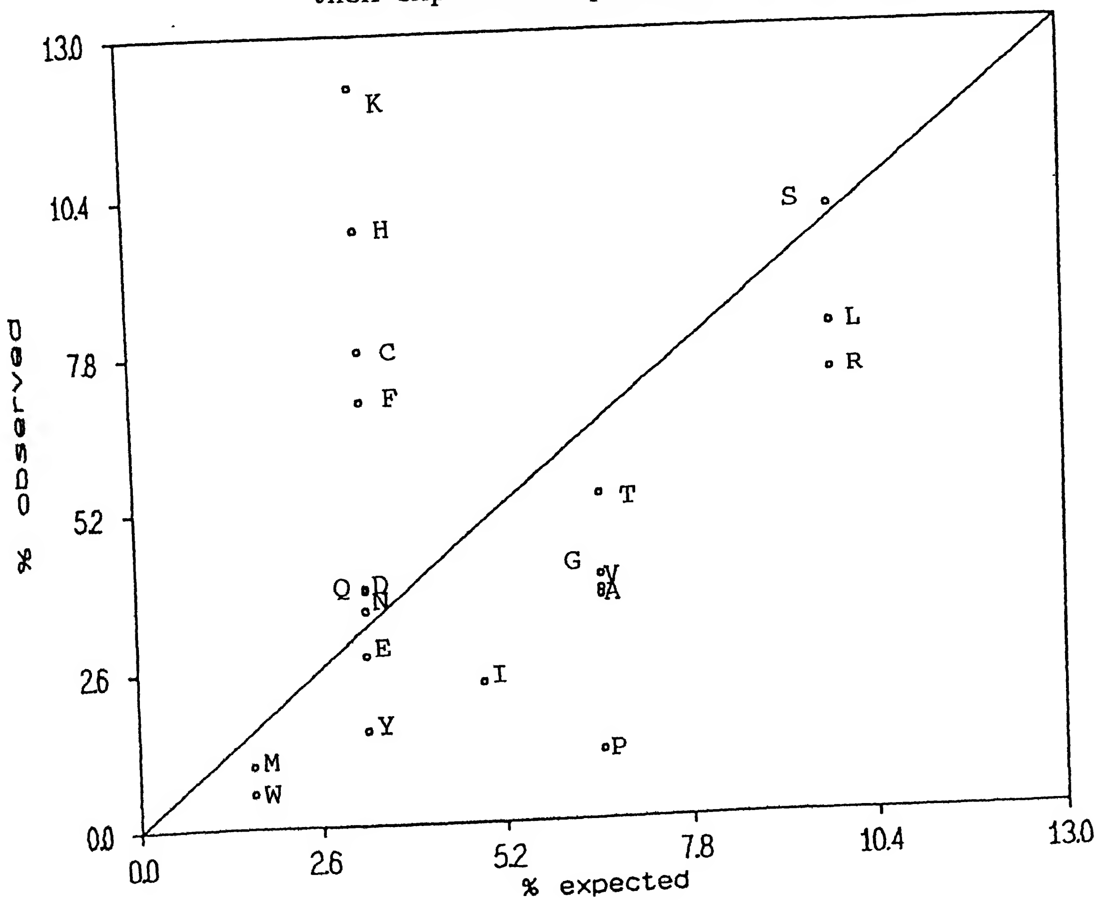
Distribution of amino acids found in zinc finger modules vs. their expected frequencies if codons occurred randomly



Amino Acid Residue	Total No.	%Observed	%Expected
Alanine	112	3.13	6.55
Cysteine	299	8.38	3.27
Aspartic Acid	114	3.19	3.27
Glutamic Acid	147	4.11	3.27
Phenylalanine	176	4.93	3.27
Glycine	171	4.79	6.55
Histidine	355	9.94	3.27
Isoleucine	120	3.36	4.91
Lysine	349	9.78	3.27
Leucine	225	6.30	9.83
Methionine	055	1.54	1.63
Asparagine	100	2.80	3.27
Proline	142	3.97	6.55
Glutamine	156	4.37	3.27
Arginine	256	7.17	9.83
Serine	270	7.56	9.83
Thereonine	246	6.89	6.55
Valine	141	3.95	6.55
Tryptophan	018	0.50	1.63
Tyrosine	116	3.25	3.27
Total	3568		

CHART.C.II.11

Distribution of amino acids found in DNA contact region of zinc finger modules vs. their expected frequencies if codons occurred randomly



Amino Acid Residue	Total No.	%Observed	%Expected
Alanine	070	3.70	6.55
Cysteine	148	7.83	3.27
Aspartic Acid	073	3.86	3.27
Glutamic Acid	053	2.80	3.27
Phenylalanine	132	6.98	3.27
Glycine	076	4.02	6.55
Histidine	186	9.84	3.27
Isoleucine	044	2.32	4.91
Lysine	230	12.16	3.27
Leucine	153	8.09	9.83
Methionine	020	1.05	1.63
Asparagine	067	3.54	3.27
Proline	022	1.16	6.55
Glutamine	074	3.91	3.27
Arginine	138	7.30	9.83
Serine	190	10.05	9.83
Thereonine	101	5.34	6.55
Valine	072	3.80	6.55
Tryptophan	011	0.58	1.63
Tyrosine	030	1.58	3.27
Total	1890		

A careful choice in the identification of each of 30 amino acid residues that comprises a typical zinc finger module is reflected in CHARTS C.II.10 and C.II.11. This aspect can be highlighted in terms of : the secondary structure of the recognition element being α-helix, the anti parallel β-sheet alignment of the arms harbouring the cysteine residues, the turn elements associated with union of the α-helix and β-sheet, the conserved Phenylalanine and Leucine residues present, the recognized contact residues which are Arginine, Lysine, Glutamine and Asparagine and the proximal placement of Arginine and Aspartic acid in the finger recognition region to accentuate the recognition contact. Thus the 12 residues contact element show an excess of α-helix makers such as Alanine, Arginine, Leucine, a depletion of residues associated with sheet formation such as Isoleucine and the near absence of Proline connected with turns, and coupled with this, significant enhancement of the recognition contact residues, particularly Lysine and Asparagine.

A profile of the zinc finger recognition region based on the above analysis is presented in CHART.C.II.12. It can be seen from this Chart that the Phenylalanine at 4 location and Leucine at 10 location are highly conserved. This criterion is important in the construction of such 12 residue modules based on general neighbourhood preferences seen in proteins. Thus Phenylalanine at 4 location and Leucine at 10 location provided two independent markers for the construction of a consensus 12 residue sequence from a general data base.

Endeavours to match the preference profile present in zinc finger recognition domain with that from the general data base were operated upon, in order to avoid extremely complex situation, using a very simple principle, namely, that any of the best likely neighbours from the general data base could be matched with any 4 best preferences pertaining to each residue of zinc finger recognition domain. It was considered that this approach was very reasonable in view of inherent anomalies in neighbour group selection and that a match if observed within these constraints would still be extremely significant in view of very large options available for a random construction of a 12 residue peptide.

CHART.C.II.12

A PROFILE OF ZINC FINGER REGION

1	2	3	4	5	6	7	8	9	10	11	12
			F						L		

% F at 4 Location in the entire set : 107/135 = 79 %

% L at 10 Location in the entire set : 109/135 = 81 %

CHART.C.II.13

The best likely neighbours (computed) on the left as well as the right from analysis of 9416 residues in major data base for each of the coded amino acids

Left Preferences	Coded Amino Acid	Right Preferences
A (85), G (73), V (63), S (62)	A	A (85), G (78), V (64), S (61)
G (29), S (22), A (21), V (19)	C	G (41), A (24), S (23), N (17)
A (56), V (50), G (44), K (38)	D	A (58), G (50), S (47), L (37)
L (39), E (37), K (36), G (35), S (35)	E	L (51), A (44), G (37), E (37)
G (36), D (32), T (31), S (30)	F	T (32), S (31), K (27), L (25)
S (93), G (80), A (78), V (53)	G	S (80), G (80), A (73), V (55), T (54), L (53)
A (27), G (24), L (22), S (21)	H	L (28), P (22), K (22), G (20)
G (48), I (34), S (33), T (30)	I	V (38), T (35), P (32), G (32)
L (54), A (52), G (48), T (43)	K	A (55), L (54), S (53), V (44)
V (58), A (55), L (54), I (54), G (53), S (51)	L	S (59), L (54), I (54), G (51), V (47)
K (15), G (15), A (15), L (13), S (13)	M	L (18), K (17), D (15), V (13)
S (38), L (33), A (30), T (30)	N	V (38), A (35), L (35), S (32), T (32)
L (42), A (32), I (32), T (32)	P	G (44), S (44), V (33), A (33)
A (36), S (33), L (30), G (29)	Q	A (34), G (34), L (23), S (22)
L (37), G (35), V (23), A (21)	R	L (30), G (29), V (28), I (21)
G (80), A (61), L (59), S (56), K (53)	S	G (93), A (62), S (56), L (51)
G (54), A (51), V (47), S (47)	T	A (49), S (45), K (43), G (43)
A (64), G (55), L (47), V (46)	V	A (63), L (58), G (53), D (50), S (49), K (44)
S (15), A (12), E (12), T (10), V (10)	W	G (16), V (15), D (13), A (13)
S (35), G (34), D (30), K (29)	Y	G (32), A (28), T (27), E (26), S (26)

CHART.C.II.14

The best likely neighbours (computed) on the left as well as the right from analysis of 3009 residues in α-helix data base for each of the coded amino acids

Left Preferences	Coded Amino Acid	Right Preferences
A(39), L(29), V(29), E(27)	A	A(39), L(32), K(29), V(24)
V(6), D(5), G(5), Y(5)	C	G(9), V(6), A(4), N(4)
A(21), E(17), V(15), L(12)	D	A(19), V(15), E(13), L(13)
E(20), A(19), L(18), K(16)	E	A(27), L(27), E(20), D(17)
A(12), E(12), D(11), N(11)	F	K(17), L(13), A(12), T(12)
A(19), L(18), S(12), H(10), V(10)	G	A(12), L(12), E(11), I(11), V(11)
A(10), L(8), S(8), K(7)	H	K(12), A(11), G(10), L(10)
L(18), A(16), G(11), I(11), R(11)	I	A(12), L(12), V(12), I(11),
A(29), L(23), F(17), K(16)	K	A(21), L(21), E(16), K(16), V(16), S (16)
A(32), E(27), V(24), L(22)	L	A(29), K(23), L(22), V(22)
L(8), E(7), A(6), K(6), M(6), S (6)	M	K(11), M(6), L(6), D(5)
L(16), V(11), A(7), T(7)	N	V(14), L(13), A(12), F(11)
I(8), L(7), T(5), N(4)	P	E(11), V(8), N(6), A(5), K(5)
A(17), E(15), L(13), Q(9), R(9)	Q	A(14), L(12), E(9), Q(9)
L(16), A(10), V(9), Q(8)	R	E(11), I(11), L(11), A(10)
L(21), K(16), A(11), V(11)	S	A(23), E(12), G(12), L(12)
A(17), L(14), F(12), V(12)	T	A(16), L(16), E(13), V(13)
A(24), L(22), K(16), D(15)	V	A(29), L(24), K(16), D(15)
E(7), A(6), Q(3), V(3)	W	L(9), A(7), G(4), S(3)
A(11), D(11), K(7), L(7), T(7)	Y	L(12), A(8), E(6), C(5), Q (5)

CHART.C.II.15

The observed occurrence of coded amino acid residues upto 4 preferences in each of the sites comprising the zinc finger recognition elements derived from data base of 135 zinc finger modules

Site	Preferences
1	G (43), D (28), E (10), Q (09)
2	K (77), R (21), Y (06), L (05)
3	S (26), K (18), R (13), A (12), G (12), T (12)
4	F (Conserved)
5	S (31), T (23), I (18), A (13)
6	Q (32), R (21), T (12), D (11), H (11), K (11)
7	K (28), R (25), S (19), N (14)
8	S (61), A (08), D (08), G (07)
9	N (19), S (18), D (16), A (16)
10	L (Conserved)
11	V (21), K (20), T (15), L (14), I (14)
12	K (43), R (26), V (10), N (09)

CHART.C.II.13 presents the best likely neighbours on the left as well as the right from analysis of 9416 residues in major data base for each of the 20 coded amino acids. A similar exercise was performed with respect to α-helix domain and presented in CHART.C.II.14. For comparison purposes the observed occurrence of coded amino acid residues upto 4 preferences in each of the sites comprising the zinc finger recognition element derived from data base of 135 zinc finger recognition region are presented in CHART.C.II.15.

The preference profile presented in CHART.C.II.13 enables the construction of 12 residue zinc finger recognition region based either on the conserved Phenylalanine at 4 location or the conserved Leucine at 10 location. Similarly, the preference profile presented in CHART.C.II.14 enables the construction of 6 residue zinc finger helical region based on Lysine at 7 location. Sequence analysis of 135 zinc finger module recognition profile has led to actual construction of a consensus sequence (CHART.C.II.16A). This has been compared with that constructed from the basic data set based on (1) the conserved Phenylalanine at 4 location (2) the conserved Leucine at 10 location and (3) the helical segment. Thus 3, 12 residue sequences and one 6 residue sequence emerged as shown in CHART.C.II.16. CHART.C.II.16A is constructed from the actual zinc finger data base, CHART.C.II.16B and CHART.C.II.16C constructed from the major data base using as markers, respectively Phe and Leu, and CHART.C.II.16D constructed from the α-helix data base. One could see from CHART.C.II.16 that the agreement here is most remarkable. The findings here provide strong support for the presence of small information-function composites, possible precursors towards evolution of present day functional and informational molecules.

C.III. A SEQUENCE ANALYSIS CORRELATION OF NON RIBOSOMAL PEPTIDES AND RIBOSOMAL PROTEINS

The notion that imprints of the early stages of protein evolution are likely to be present in present day functional proteins was proved in the earlier section by comparison of the modular Zn-finger proteins with data base sequences. The significant agreement that was found here that tended to agree with a modular basis for protein evolution made it logical to examine the correlation if any, pertaining to neighbour preferences present in non ribosomal peptides with the cytoplasmic biosynthesis of gramicidin and tyrocidine, both in terms of small peptide blocks that are aggrandized by specific protein ensemble as well as the linking site leading to specific peptide bond formation.

The alternate method for selective peptide bond formation not involving the information system and which takes place in the cytoplasmic domain of cellular system was recognized in the fifties and because of their clinical application as antibiotics has received considerable attention. As stated in SECTION.B the relationship between fatty acid biosynthesis and that of biosynthesis of non ribosomal peptides was recognized early. However, the protocols pertaining to the latter are necessarily more complicated because of the need to construct the peptide in a highly selective manner. This is accomplished with the help of highly complex multienzymatic systems, which have a dual function, namely, the selective construction of blocks of small peptides and to promote the interlinking of such molecules. Thus there is a superficial similarity between the construction of Zn-finger proteins and non ribosomal peptides. Individual enzyme ensembles have been shown under cell free environments to promote the construction of peptides from amino acid precursors in a highly selective manner. In this context, it was considered of interest to determine whether a generalized neighbour preference approach is associated with the construction of small peptide molecules by enzyme complexes. This aspect was examined via protocols similar to that adapted and presented in SECTION.C.I.

Endeavours to secure data base pertaining to sequences in peptides of non ribosomal

origin did not succeed, consequently the construction of such a data base had to be performed manually. In TABLE.C.III.1 is presented 73 of non ribosomal peptides which have been used in the construction of data base pertaining to the present analysis. Unlike proteins of ribosomal origin a rational analysis here is complicated by the intervention of non coded amino acids and other structural units within the peptide frame work. This aspect has been clearly shown in TABLE.C.III.1. Necessarily, therefore the sequence analysis had to be restricted to amino acids that are present in the code complement including their D - analogs.

Analysis of such a data base comprising of 482 residues which also indicate the "Left-Right" preference is shown in CHART.C.III.1. A comparison of CHART.C.III.1 with the corresponding one pertaining to amino acids of ribosomal origin do tend to show that both neighbour preferences as well as "Left-Right" preferences pertaining to a particular residue is somewhat accentuated here. It nevertheless provides a starting point for the analysis of biosynthetic pathways associated with non ribosomal peptide synthesis.

The established pathway for tyrocidine biosynthesis[14] is illustrated in CHART.C.III.2. This cyclic peptide having 10 residues is constructed by a three enzyme ensembles. The biosynthesis is initiated with a peptide bond formation involving D-Phe harboured by light enzyme which is transformed to an intermediate enzyme which is able to construct and hold the module Pro-Phe-D-Phe. The tetrapeptide thus formed is then transferred from the intermediate enzyme to heavy enzyme involving the D-Phe-Asn peptide bond. Perhaps the most complex system involved in tyrocidine biosynthesis, namely, heavy enzyme, assembles the penta peptide Asn-Gln-Tyr-Val-Orn-Leu and accepts the tetrapeptide module constructed by earlier system to provide the open deca-peptide. The biosynthetic cycle is then completed with peptide bond formation involving Leu (residue 10) and D-Phe (residue 1).

The total biosynthetic cycle can be analyzed in terms of the construction of the module Pro-Phe-D-Phe by the intermediate enzyme and Asn-Gln-Val-Tyr-Orn-Leu by

TABLE.C.III.1

Listing of non ribosomal peptides used in the construction of the data base

No.	Name	α-Amino Acid (AA) Residues: Coded AA	D-AA	Others
1.	Antamanide	10	-	-
2.	Anaphylatoxin	08	-	-
3.	Bacitracin-F	09	-	01
4.	Bacillomycin L	04	03	01
5.	Bacitracin A	07	02	03
6.	Cycloamanide-I	07	-	-
7.	Cycloamanide-II	08	-	-
8.	Cycloamanide-III	08	-	-
9.	Bottromycin A_2	03	-	-
10.	BE-4	04	-	01
11.	Bacillus Subtilis C-756 metabolite	07	-	-
12.	Cyclolinopeptide	09	-	-
13.	Desthiomalformin	02	03	-
14.	Deoxybouvardin	06	-	-
15.	Didemnin-A	05	-	-
16.	Evolidine	07	-	-
17.	Enkephalin	05	-	-
18.	Echinocandin-D	05	-	-
19.	Fungisporin	04	04	-
20.	Gramicidin-A	09	06	-
21.	Gramicidin-SA	06	02	02
22.	Gramicidin-J_1	03	02	02

No.	Name	Coded AA	D-AA	Others
23.	Gramicidin-J_2	02	02	02
24.	Griselimycin	06	-	04
25.	Gratisin	10	-	02
26.	Leupeptin	03	-	-
27.	Longicatenamycin	02	02	02
28.	Lophyrotomin	04	04	-
29.	Malformin A	02	03	-
30.	Malformin C	01	04	-
31.	Mulndocandin	04	-	-
32.	Mycobacillin	07	06	-
33.	Mycosubtilin	04	04	-
34.	Norsufactin	07	-	-
35.	Polymyxin E	07	01	-
36.	Polymyxin M	07	01	-
37.	Peptidoglycan	02	02	01
38.	Phalloin	07	-	-
39.	Retroantamanide	10	-	-
40.	Retrogramicidin S	06	02	02
41.	Tyrocidin A	07	02	01
42.	Tyrocidin B	07	02	01
43.	Tyrocidin C	07	02	01
44.	Tyrocidin E	07	02	01
45.	Tuberactinomycin	08	02	-
46.	Terlipressin	12	-	-
47.	TL-119	06	-	01
48.	Rhozonin A	05	-	02

No.	Name	Coded AA	D-AA	Others
49.	Rhozonin B	05	-	02
50.	Viscunamide	02	03	-
51.	HC-Toxin	01	02	-
52.	Viscosin	05	-	-
53.	Gramicidin S	06	02	02
54.	Actinomycin-C_2	04	02	04
55.	Aniso-Actinomycin D	04	02	04
56.	Actinomycin.X	03	02	05
57.	Esperin	04	01	-
58.	Geodiamolides	03	-	-
59.	Glumamycin	07	-	02
60.	Lanthionine	14	-	-
61.	Mycoplanecin A	09	-	-
62.	Norphalloin	07	-	-
63.	Nummularia	03	-	-
64.	Protodestruxin	04	-	01
65.	Isarin	04	01	-
66.	Sativanine E	03	-	-
67.	Triostin A	08	-	-
68.	Vernamycin B	04	-	01
69.	Dorcidin	05	-	01
70.	Ostreogrycin B	04	-	01
71.	K-13	03	-	-
72.	Beavellide	03	-	-
73	Dextruxin	03	-	-

CHART.C.III.1

MASTER ANALYSIS FOR NEIGHBOUR (LEFT-RIGHT) PREFERENCE

(Non Ribosomal Peptides, 482 residues)

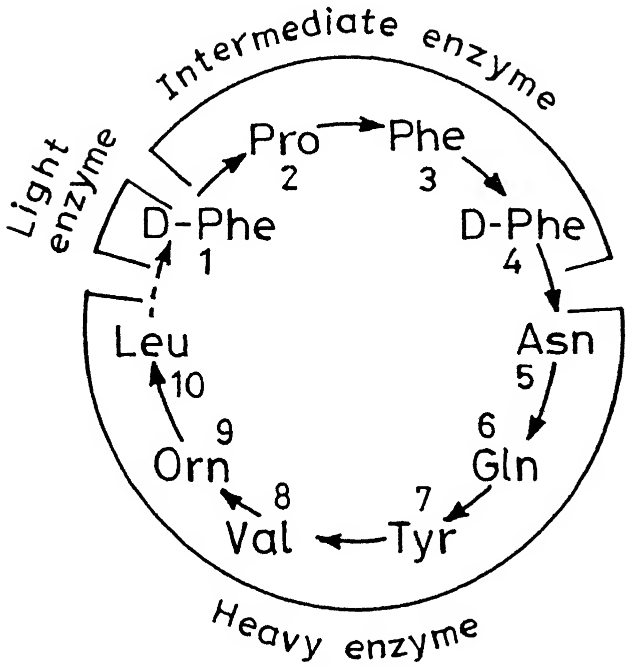
NEIGHBOUR	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	05/05	03/00	01/02	00/00	02/02	01/02	00/00	01/01	00/01	02/06	00/00	00/00	04/03	00/00	01/00	01/02	01/03	03/03	02/00	03/03
C	00/03	01/01	00/00	00/02	02/00	01/02	00/00	00/01	00/00	02/00	00/00	00/01	03/00	00/00	00/00	00/00	00/01	04/01	00/00	01/00
D	02/01	00/00	03/03	02/02	00/01	02/01	00/02	00/01	00/00	03/01	00/00	00/00	01/01	00/00	00/00	01/02	00/00	01/02	00/00	02/02
E	00/00	02/00	02/02	00/00	00/00	00/00	00/02	02/00	00/00	02/03	00/00	00/00	00/00	01/00	00/00	01/01	00/00	00/00	00/00	01/00
F	02/02	00/02	01/00	00/00	11/11	02/03	02/00	00/02	00/00	08/11	00/00	03/00	15/10	01/01	00/00	01/02	01/00	05/05	00/00	00/02
G	02/01	02/01	01/02	00/00	03/02	03/03	02/00	00/00	00/01	02/05	00/00	00/00	01/00	00/00	00/00	01/01	00/00	01/01	00/00	00/00
H	00/00	00/00	02/00	02/00	00/02	00/02	00/00	00/00	00/00	01/00	00/00	00/00	00/00	00/00	00/00	00/01	00/00	00/00	00/00	00/00
I	01/01	01/00	01/00	00/02	02/00	00/00	00/00	01/01	03/00	03/05	00/00	00/00	01/04	00/00	00/00	00/00	00/00	03/01	00/00	00/00
K	01/00	00/00	00/00	00/00	00/00	01/00	00/00	03/00	02/02	01/01	00/00	00/00	00/01	00/00	00/00	00/00	05/05	00/00	00/00	00/00
L	06/02	00/02	01/03	03/02	11/08	05/02	00/01	05/03	01/01	08/08	00/02	00/01	07/04	00/00	01/00	01/00	01/01	04/05	03/04	00/00
M	00/00	00/00	00/00	00/00	00/00	00/00	00/00	00/00	00/00	02/00	00/00	00/00	00/02	00/00	00/00	00/00	00/00	00/00	00/00	00/00
N	00/00	01/00	00/00	00/00	00/03	00/00	00/00	00/00	00/00	01/00	00/00	01/01	00/00	05/01	00/00	01/00	00/00	00/01	00/01	00/01
P	03/04	00/03	01/01	00/00	10/15	00/01	00/00	04/01	01/00	04/07	02/00	00/00	05/05	00/01	00/00	01/01	03/02	07/10	01/02	05/00
Q	00/00	00/00	00/00	00/01	01/01	00/00	00/00	00/00	00/00	00/00	00/00	01/05	01/00	00/00	00/00	00/00	00/00	00/00	00/00	03/00
R	00/01	00/00	00/00	00/00	00/00	00/00	00/00	00/00	00/00	00/01	00/00	00/00	00/00	00/00	00/00	00/00	00/00	00/00	00/00	00/00
S	02/01	00/00	02/01	01/01	02/01	01/01	01/00	00/00	00/00	00/01	00/00	00/01	01/01	00/00	00/00	00/00	01/00	01/01	00/00	00/02
T	03/01	01/00	00/00	00/00	00/01	00/00	00/00	00/00	05/05	01/01	00/00	00/00	02/03	00/00	00/00	00/01	00/00	05/01	00/00	00/00
V	03/03	01/04	02/01	00/00	05/05	01/01	00/00	01/03	00/00	05/04	00/00	01/00	10/07	00/00	00/00	01/01	01/05	07/07	01/01	00/05
W	00/02	00/00	00/00	00/00	00/00	00/00	00/00	00/00	00/00	04/03	00/00	01/00	02/01	00/00	00/00	00/00	00/00	01/01	00/00	00/01
Y	03/03	00/01	02/02	00/01	02/00	00/00	00/00	00/00	00/00	00/00	00/00	01/00	00/05	00/03	00/00	02/00	00/00	05/00	01/00	03/03

CENTRAL RESIDUE

Read out format: Central residue X can have neighbour profile YX(left preference) or XY(right preference) example for P, FP:PF :: 15:10

CHART.C.IV.2.

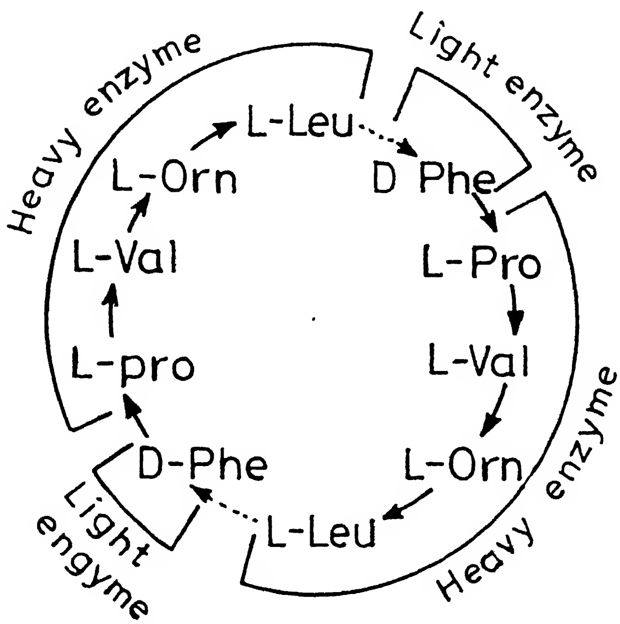
Tyrocidine biosynthesis: correlation of sequences of the module involved and their linker residue with the data base.



MODULE	CONSTRUCTED FROM DATA BASE
Asn-Gln-Tyr-Val-Orn-Leu	Asn-Gln-Tyr-Val-X-Leu
LINKER RESIDUES	PREDICTED FROM DATA BASE
Phe → Pro	Phe → Pro (FP/PF : 15/10)
Phe → Asn	not predictable (FN/NF : 0/0)
Leu → Phe	Leu → Phe (LP/PL : 7/4)

CHART.C.IV.3

Gramicidin biosynthesis : correlation of sequences of the module involved and their linker residue with the data base.



MODULE	CONSTRUCTED FROM DATA BASE
Pro-Val-Orn-Leu	Pro-Val-X-Leu

LINKER RESIDUES	PREDICTED FROM DATA BASE
Leu → Phe	Leu → Phe (LP/PL : 7/4)
Phe → Pro	Phe → Pro (FP/PF : 15/10)

heavy enzyme and peptide bond formation at 3 sites leading to D-Phe-Pro, D-Phe-Asn and Leu-D-Phe. From the above it is clear that the grand enzyme assembly involved in tyrocidine biosynthesis reflects aspects of a high degree of evolution. In the construction of 2 individual modules, the constituent amino acids are precisely aligned by the enzyme complex for sequence specific peptide bond formation. In terms of optimization of genetic information it is reasonable to assume that each of the enzyme subsystems charged with harbouring an individual amino acids must have common features across the non ribosomal peptide biosynthesis. In view of the high degree of evolution represented here the construction of the module must have a bearing on preferences for neighbours which being largely dependent on the nature of the side chain ought to show common features involved in the construction of similar modular units, presently directed by information system (SECTION.C.I.). This line of argument is also applicable with respect to the inter enzyme ensemble peptidation. Thus a comparison of the sequences present in the module and the alignment of the ensemble in terms of peptidation would be of interest.

Various stages in tyrocidine biosynthesis are compared with those constructed from data base (CHART.C.III.2). It could be seen from this that the comparison here is very good which tends to support the notion that preferences in the choice of neighbours of the α-amino acid as well as their placement with respect to peptide bond formation must have played a key role in the evolution of enzyme ensemble associated with tyrocidine biosynthesis.

A similar exercise was performed with respect to gramicidin and illustrated in CHART. C.III.3.

The correlations observed above are truly significant since they not only highlight the advantages inherent in construction of polypeptides by a blockwise method but also show that such processes are perfected with the help of enzymes for error free genesis of the compounds.

C.IV. CHEMOSELECTIVITY AND PROTEIN SECONDARY STRUCTURE : THE GRAMICIDIN PORE - SELECTIVE TRYPTOPHAN REPLACEMENT WITH ASPARTIC ACID

In SECTION.C.I. an analysis pertaining to the neighbour selection as a function of secondary structures namely, α-helix and β-sheet were made (CHART.C.I.8 and CHART. C.I.9). It is obvious that the placement of a specific amino acid residue is associated with the generation of a unique set of environment that is needed to bring about the specific function pertaining to the protein. A corollary to this would be the notion that the chemical reactivity profile of a particular coded amino acid side chain would vary as a function of its placement in the protein manifold. The organic chemistry of the 20 coded amino acid side chains within the protein manifold, that play such an important role in protein structure, protein folding, protein design and protein function, has not received its due attention. Since the pioneering endeavours of Witkop in the fifties developments here have not been a sustained one. Experimental difficulties and skepticism relating to our ability to perform selective operations within the confines of highly intricate protein structures continue to haunt this domain. Yet, the delineation of chemoselectivity, arising from protein structural constraints will provide practical inputs across the entire domain and thus merits attention. Recent pioneering studies have shown that such selective operations are possible and that they open up avenues for practical application[26–31]. The development of, during the past decade in our laboratory, methodologies for coded amino acid side chain modification, the achievement of target side chain selectivity over competing sites and finally the delineation of preferences among the same side chain made it logical to experimentally demonstrate the manifestation of such chemoselectivity.

At the outset conditions are defined as to the choice of substrate to be used for this demonstration. These are (i) the number of residues should be relatively small (ii) the target side chain must be present within the peptide frame work at several locations (iii)

the structure of the peptide to be known in terms of x-ray crystallography (iv) competing residues should be absent and (v) the substrate preferentially should have interesting biological properties so that the product would be of interest. An examination of peptide sequences from this vantage enables to identify Gramicidin as the ideal target molecule.

The primary sequence, HCO-ValGlyAlaLeu•AlaVal•ValVal•TrpLeu• TrpLeu•TrpLeu• TrpNHCH$_2$ CH$_2$OH, of gramicidin A (GA), produced by Bacillus brevis, during transition from vegetative to sporulation phase, constitutes a unique illustration of the best in peptide design, wherein a relatively small peptide (molecular weight $\sim$ 1880), of non ribosomal origin, harbors latent facets, to generate by dimer formation, either double helical pores or single channels, having a hydrophilic interior and a hydrophobic surface, the former associated with sequestering and transport of mono valent ions and the latter, with chirally alternating side chains extending outwards on the same side and where the indole rings of tryptophan, by stacking and inter digitation, promote the formation of helical clusters of pores or by protrusion provide splines enabling the nestling of lipids to form stable channels. The dynamic processes associated with pore and channel formation and the transformation of the former to the latter has been well studied. In solvents of low polarity, ranging from dioxan to methanol GA and simple analogs exist almost exclusively as double helical dimers, deriving stabilization from as many as 28 intra molecular hydrogen bonds. The predominant species here is an anti parallel left handed double helix having 5.6 residues per turn, which crystallizes out from methanol in the monoclinic form and whose structure has been established by x-ray crystallography (CHART.C.IV.1). This representation brings out the unique stacking profile of Trp^9 and Trp^{11} , which also was shown to play a role in the formation of helical clusters by interdigitation [32–36].

Clearly, Trp^9 and Trp^{11} are located in the deep hydrophobic region of (1) (CHART. C.IV.3) and this aspect can be taken advantage of to bring about chemoselective changes. Thus a reagent whose affinity lies towards this region, can be expected to selectively

interact with these residues. This notion has been experimentally verified.

Gramicidin A (GA) was transformed to O-acetyl gramicidin A (GA-OAc) in 75% yields[37]. A suspension of GA-OAc (0.036 mmol) in CCl_4:CH_3CN:H_2O (1.5:1.5:3 ml) was admixed with $NaIO_4$ (0.648 mmol; 0.25 eq), $RuCl_3.3H_2O$ (< 1 mg), cooled in ice, sealed, left shaken at rt for 8 h, cooled, cautiously opened, filtered, residue washed with CCl_4 (2 ml), CH_3CN (2 ml), evaporated in vacuo, extracted with MeOH (3 x 5 ml) and evaporated. HPLC performed on the residue (0.038 g) on a reverse phase column using a dual solvent gradient system [solvent A: 0.1% TFA in water; solvent B: 0.1% TFA in acetonitrile; t(min) (A:B %): 0(100:0), 5(80:20), 45(70:30), 50(0:100)], led to the isolation of eluents of the two major peaks at retention times, respectively, 25 and 29 minutes (fractions II and I respectively).

Amino acid analysis of fraction I, at once showed the transformation of one of the four tryptophan residues to aspartic acid. Peptide sequencing showed that it was Asp^9GA-OAc (2) (CHART.C.IV.3).

The more polar fraction II, when similarly processed, was found to have the sequence Asp^9Asp^{11}GA-OAc (3) (CHART.C.IV.3).

The yields of (2) and (3) are estimated as 10% and 28% respectively, which is remarkable, since, under the conditions employed, the maximum yields of mono and di asp gramicidin are, respectively, 25% and 50%.

Parallel oxidation of GA-OAc with 4 equivalents of the reagent afforded poor yields of crude product, whose HPLC profile was quite complex.

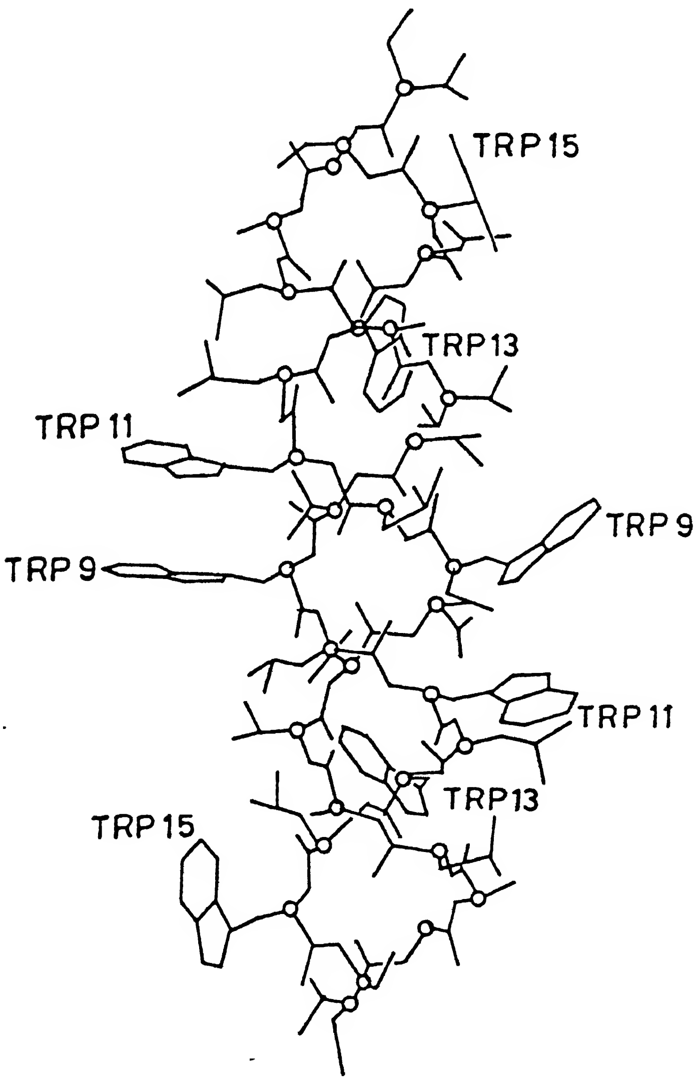
A surprising number of criteria had to be met in the transformation of (1) to (2) and (3). The Ru(VIII) mediated Trp → Asp change has been shown to proceed via sequence, N-formyl kynurinane → kynurinane → γ-oxo glutamic acid → asp[38,39]. In order to achieve selectivity, this cascade should effectively compete over attack on a fresh Trp site of the 4 trp residues located at 9, 11, 13 and 15 positions in GA, the former pair are at the hydrophobic core. Indeed crystallographic studies tend to show that

CHART.C.IV.1

Gramicidin A :

HCO-Val-Gly-Ala-Leu•-Ala-Val•-Val-Val•-Trp-Leu•-Trp-Leu•-Trp-Leu•-Trp-$NHCH_2CH_2OH$

Sequence



X-ray Structure of Gramicidin A (Pore)

CHART.C.IV.2

The Chemoselective transformation of Acetyl Gramicidin A to 9-Asp and 9-Asp, 11-Asp Gramicidin :

HCO-Val-Gly-Ala-Leu•-Ala-Val•-Val-Val•-Trp-Leu•-Trp-Leu•-Trp-Leu•-Trp-$NHCH_2CH_2OH$

$\xrightarrow{\text{(i)}}$ HCO-Val-Gly-Ala-Leu•-Ala-Val•-Val-Val•-Trp-Leu•-Trp-Leu•-Trp-Leu•-Trp-NH-CH_2CH_2OAc

$\xrightarrow{\text{(ii)}}$ HCO-Val-Gly-Ala-Leu•-Ala-Val•-Val-Val•-Asp-Leu•-Trp-Leu•-Trp-Leu•-Trp-NH-CH_2CH_2OAc (10%)

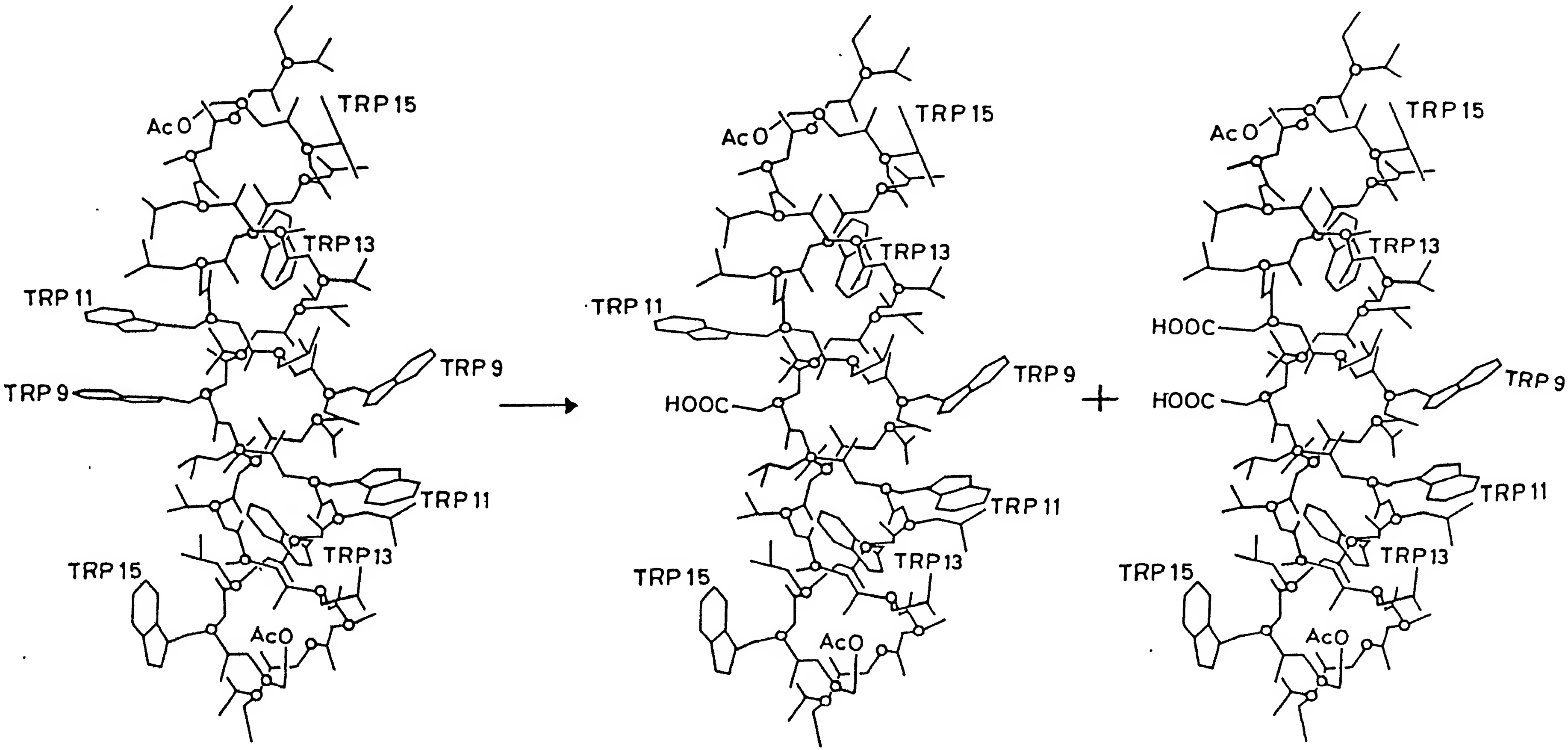
+

HCO-Val-Gly-Ala-Leu•-Ala-Val•-Val-Val•-Asp-Leu•-Asp-Leu•-Trp-Leu•-Trp-NH-CH_2CH_2OAc (28%)

(i) Ac_2O/Py

(ii) $RuCl_3.2H_2O$ / $NaIO_4$, $CCl_4:CH_3CN:H_2O$ :: 1:1:1.5 mL

CHART.C.IV.3



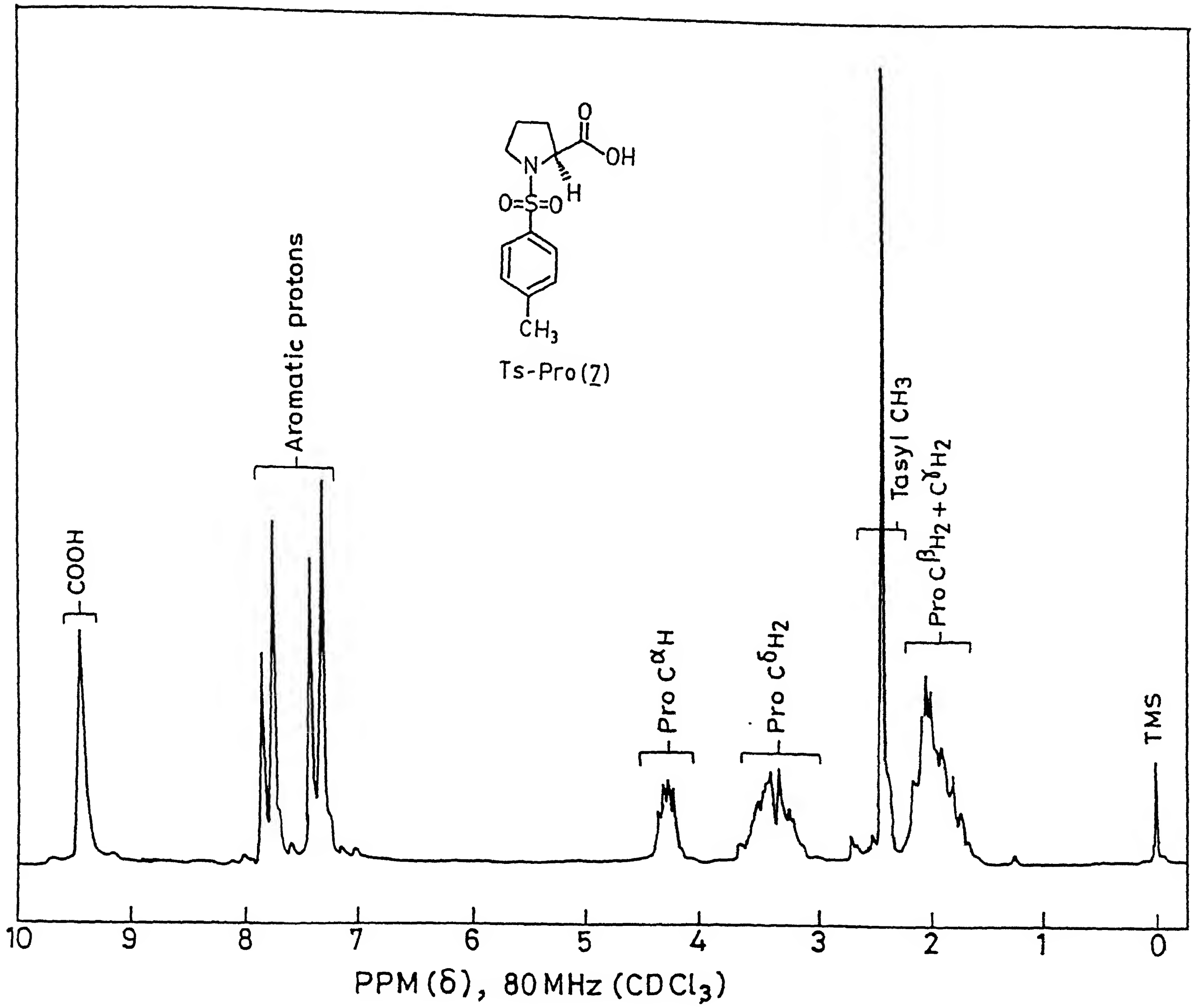
(1) (2) (3)

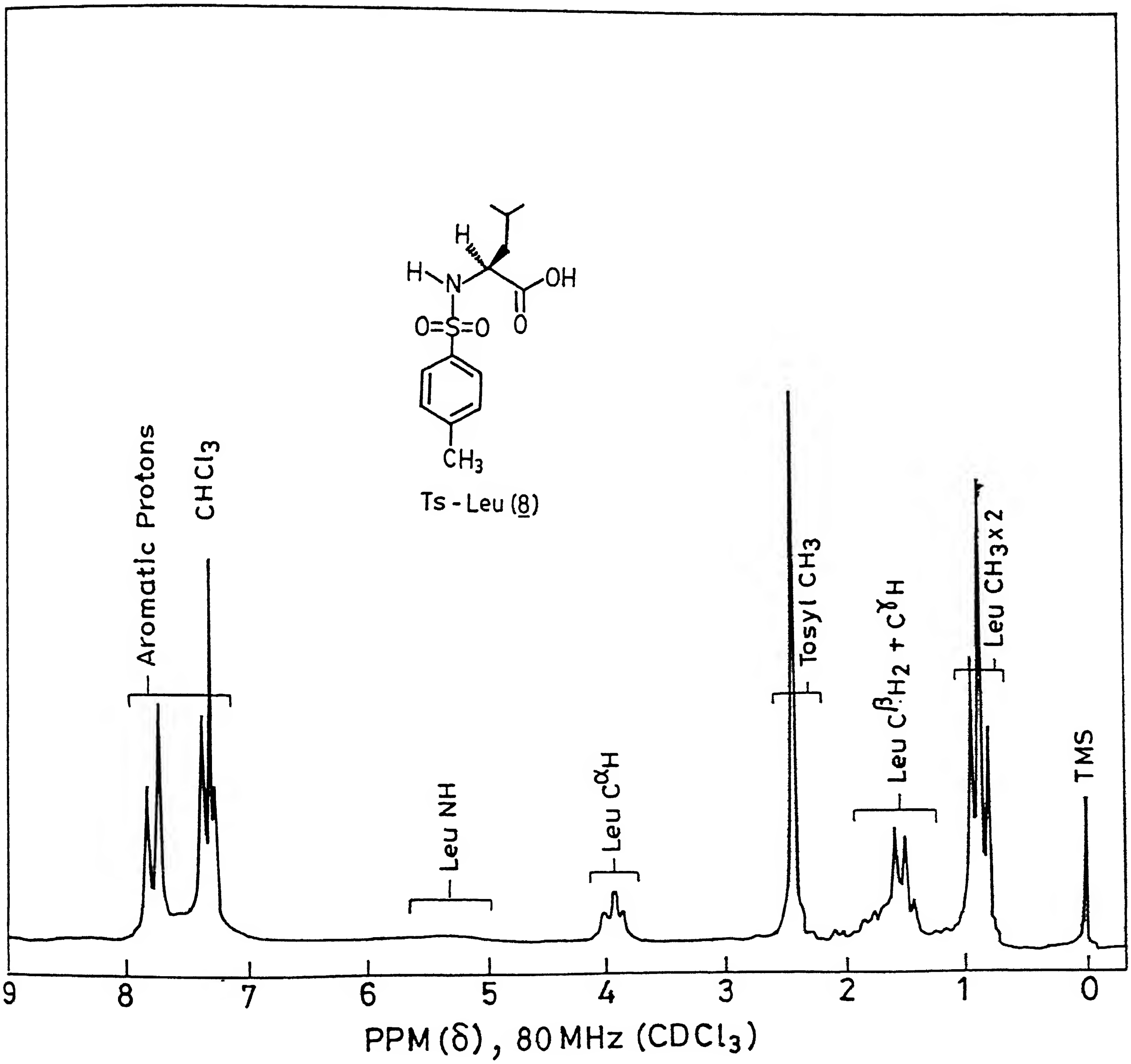
A structural profile of chemoselective Trp to Asp transformation in Gramicidin A
(for clarity changes are shown only in one of the strands)

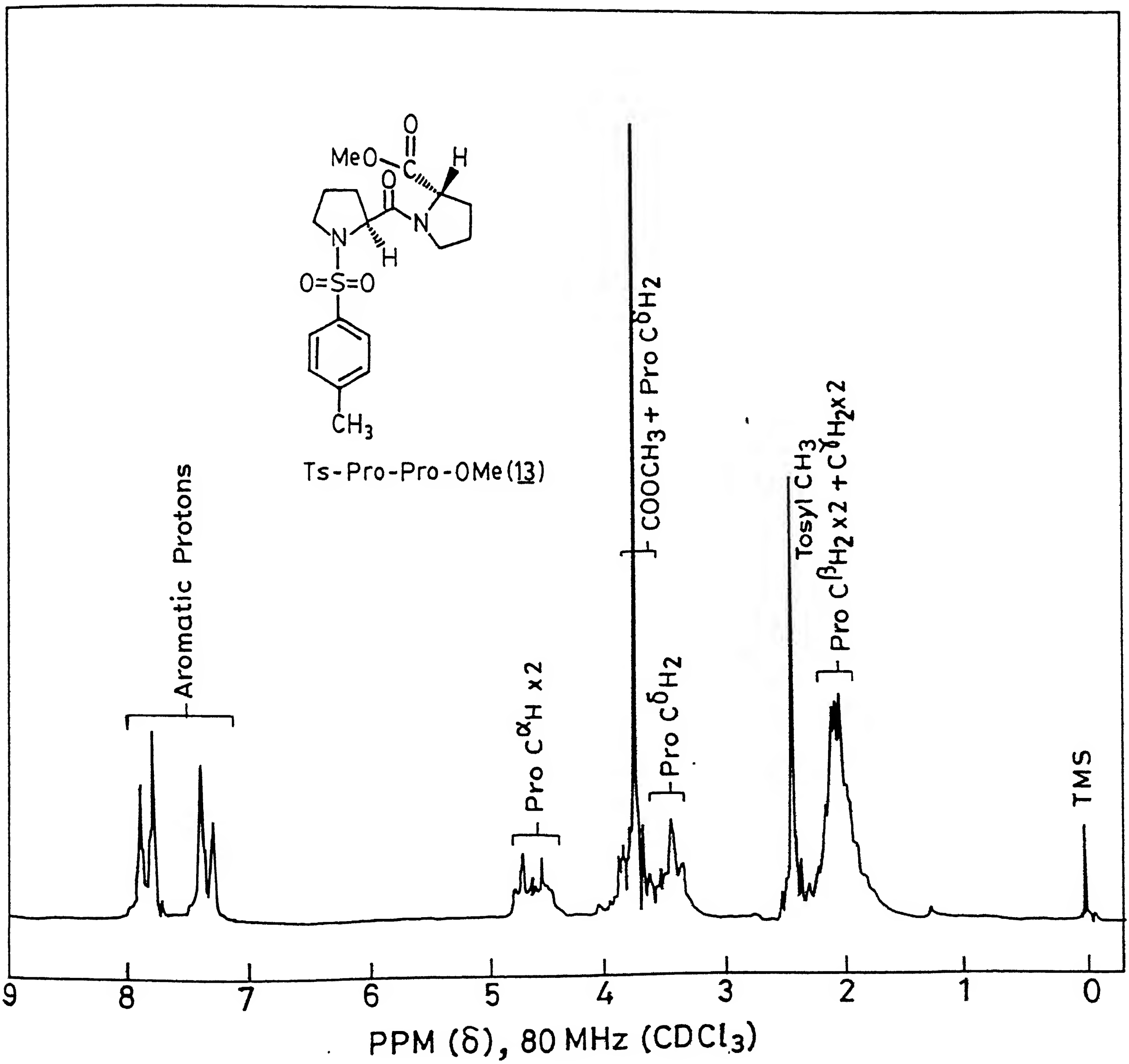
polar molecules can reach and envelope Trp^{13} residues[34]. Since in Ru(VIII) oxidations in bi-phasic media, the reagent is partitioned heavily in favour of the organic phase, the observed preference for Trp^9 and Trp^{11} is understandable. The exposed profile of Trp^9 over Trp^{11} clearly seen in (1), would explain the maximum shown for Trp^9 oxidation.

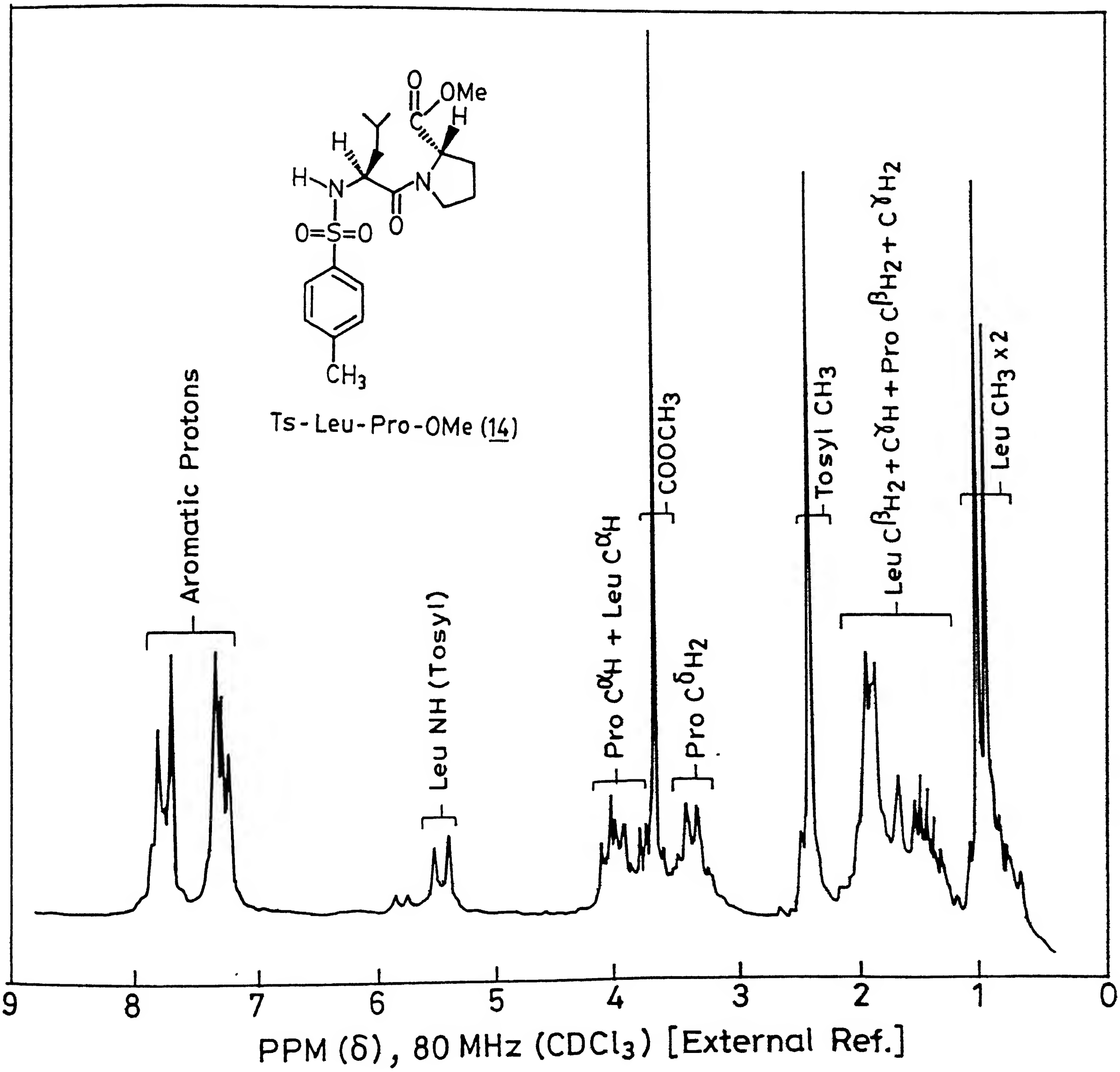
From a practical vantage, compounds (2) and (3) are excellent substrates, particularly pertaining to possibilities for the formation of pore clusters with bi-valent cations. Since the procedure presented here offers a simple route to (2) and (3) from the commercially available gramicidin, diverse studies relating to the understanding of the profile of these are planned.

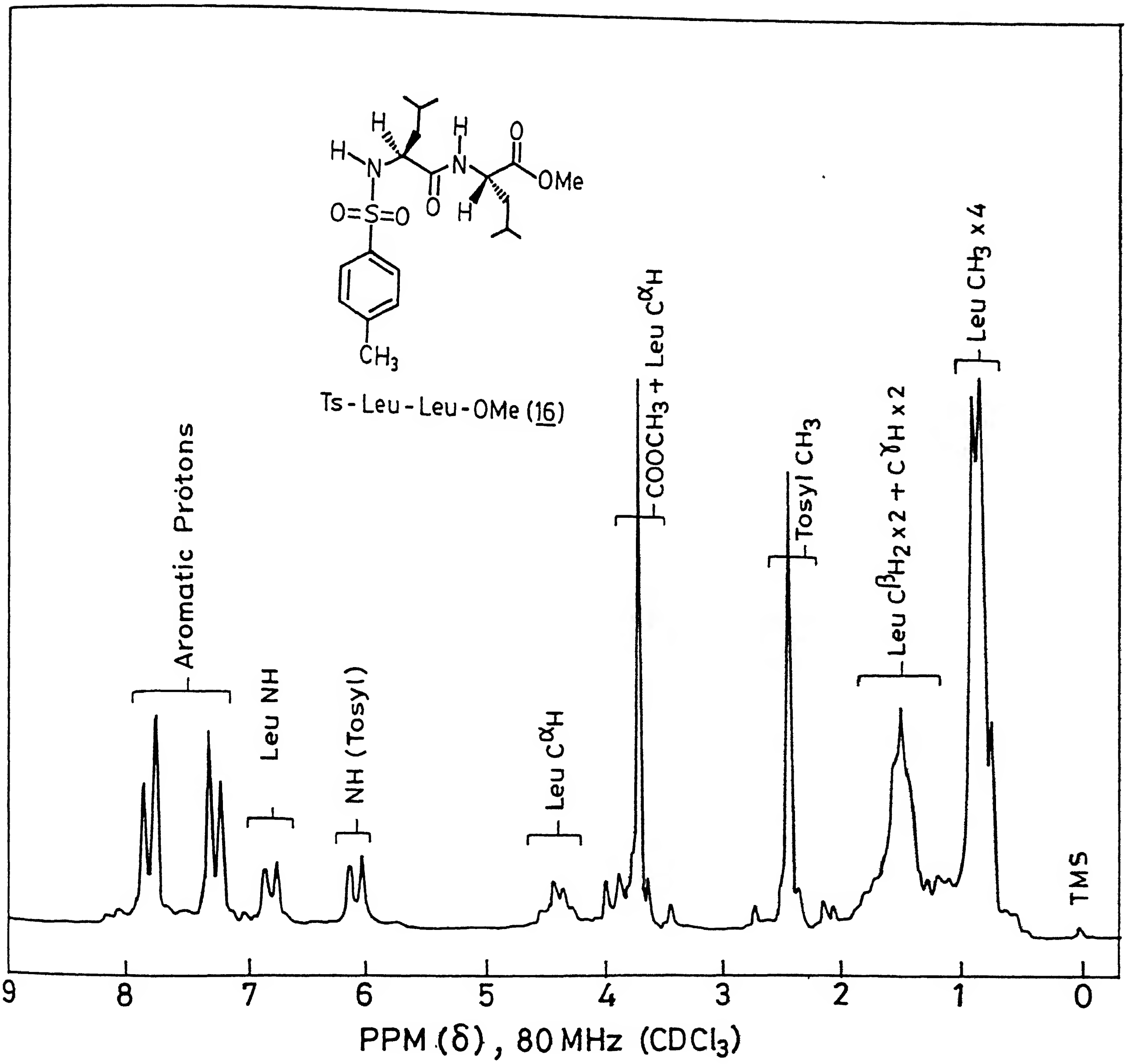
D. SPECTRA

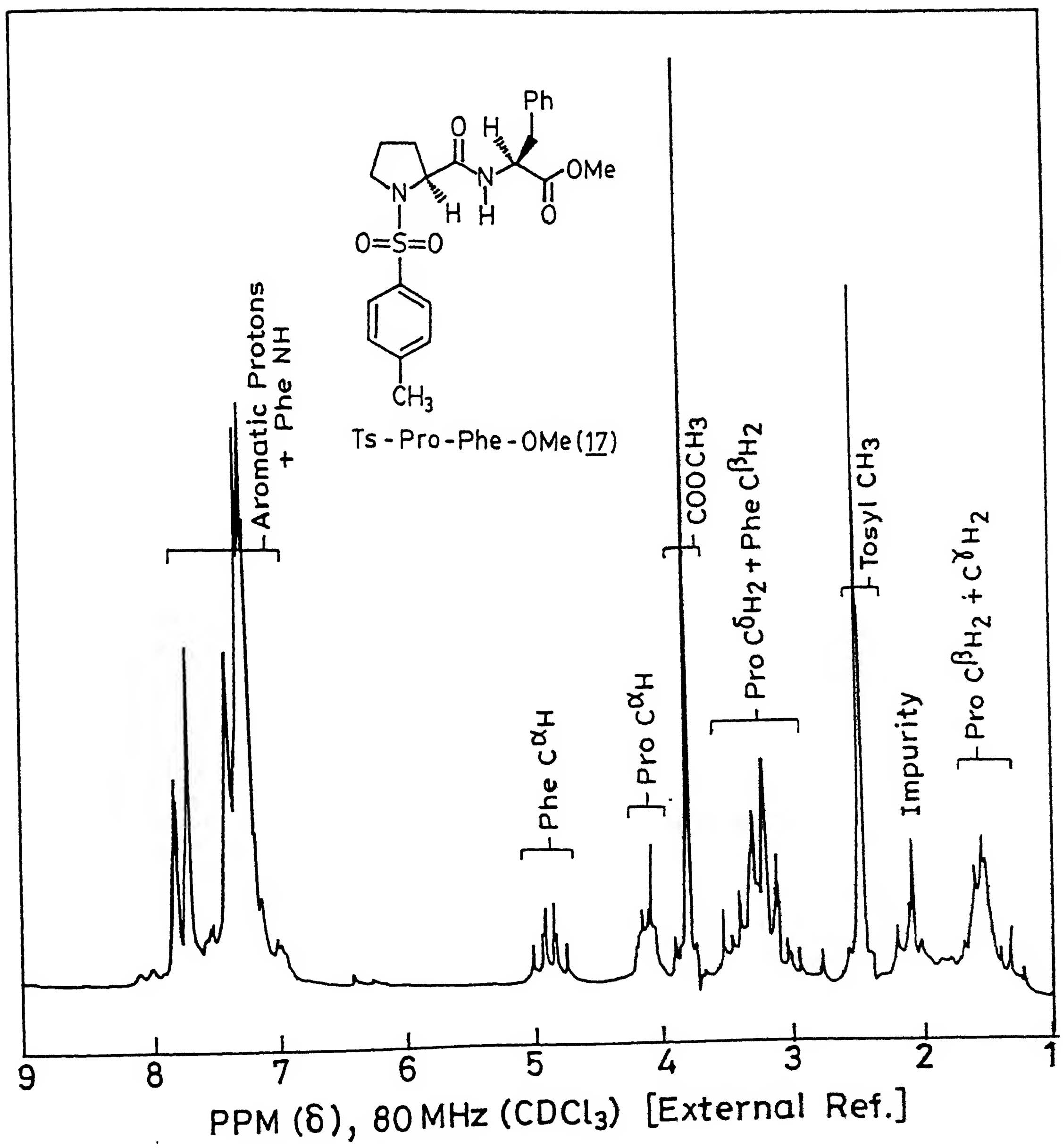

Ts-Pro(7)
COOH
Aromatic protons
Pro $C^{\alpha}H$
Pro $C^{\delta}H_2$
Tasyl CH_3
Pro $C^{\beta}H_2 + C^{\gamma}H_2$
TMS
10 9 8 7 6 5 4 3 2 1 0
PPM (δ), 80 MHz ($CDCl_3$)

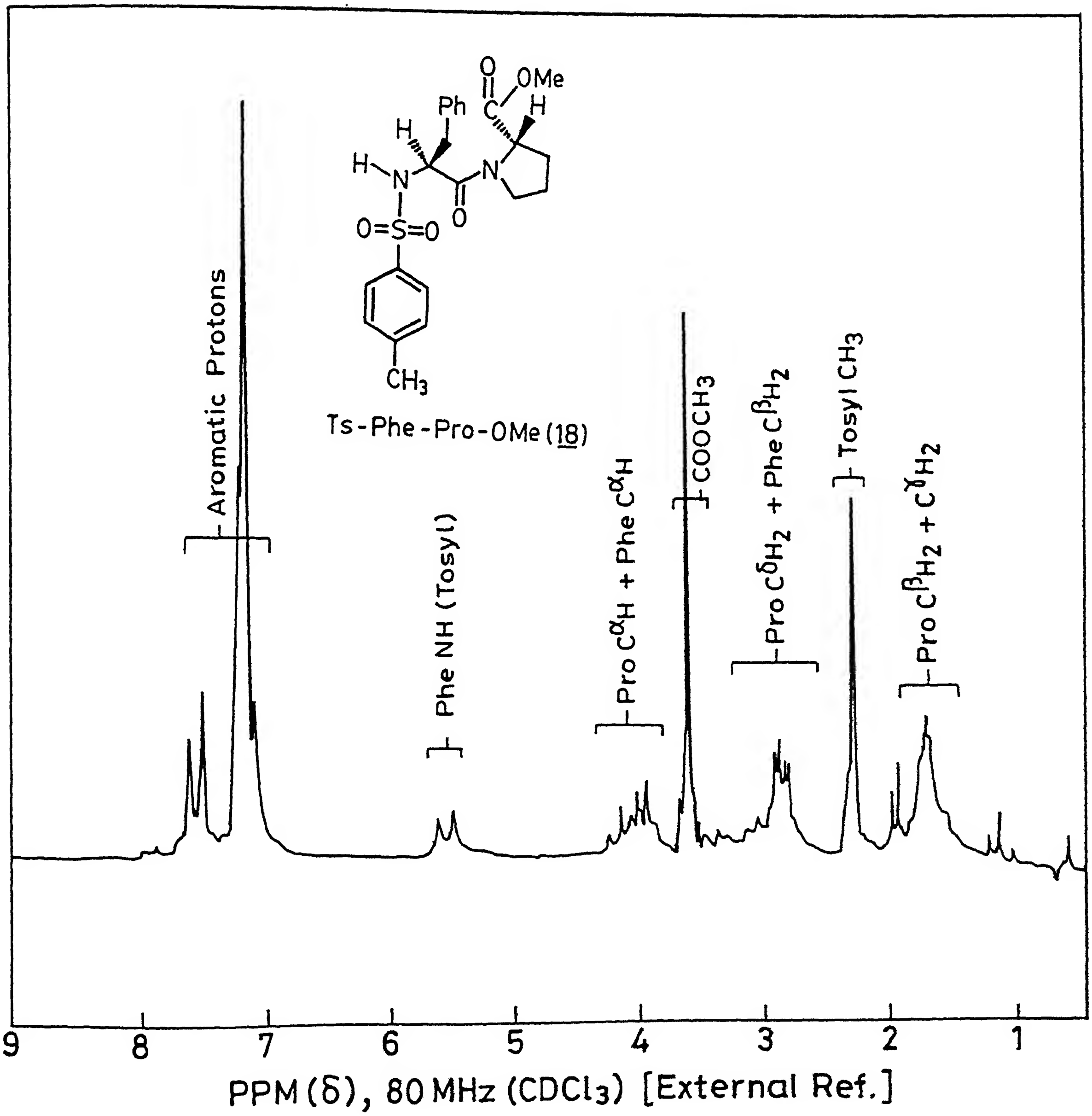

Ts - Leu (8)
Aromatic Protons
CHCl3
Leu NH
Leu CαH
Tosyl CH3
Leu CβH2 + CγH
Leu CH3 x 2
TMS
9 8 7 6 5 4 3 2 1 0
PPM (δ), 80 MHz (CDCl3)


MeO
O
H
N
O=S=O
CH3
Ts-Pro-Pro-OMe (13)
Aromatic Protons
Pro CαH x2
COOCH3 + Pro CδH2
Pro CδH2
Tosyl CH3
Pro CβH2 x2 + CγH2 x2
TMS
9
8
7
6
5
4
3
2
1
0
PPM (δ), 80 MHz (CDCl3)

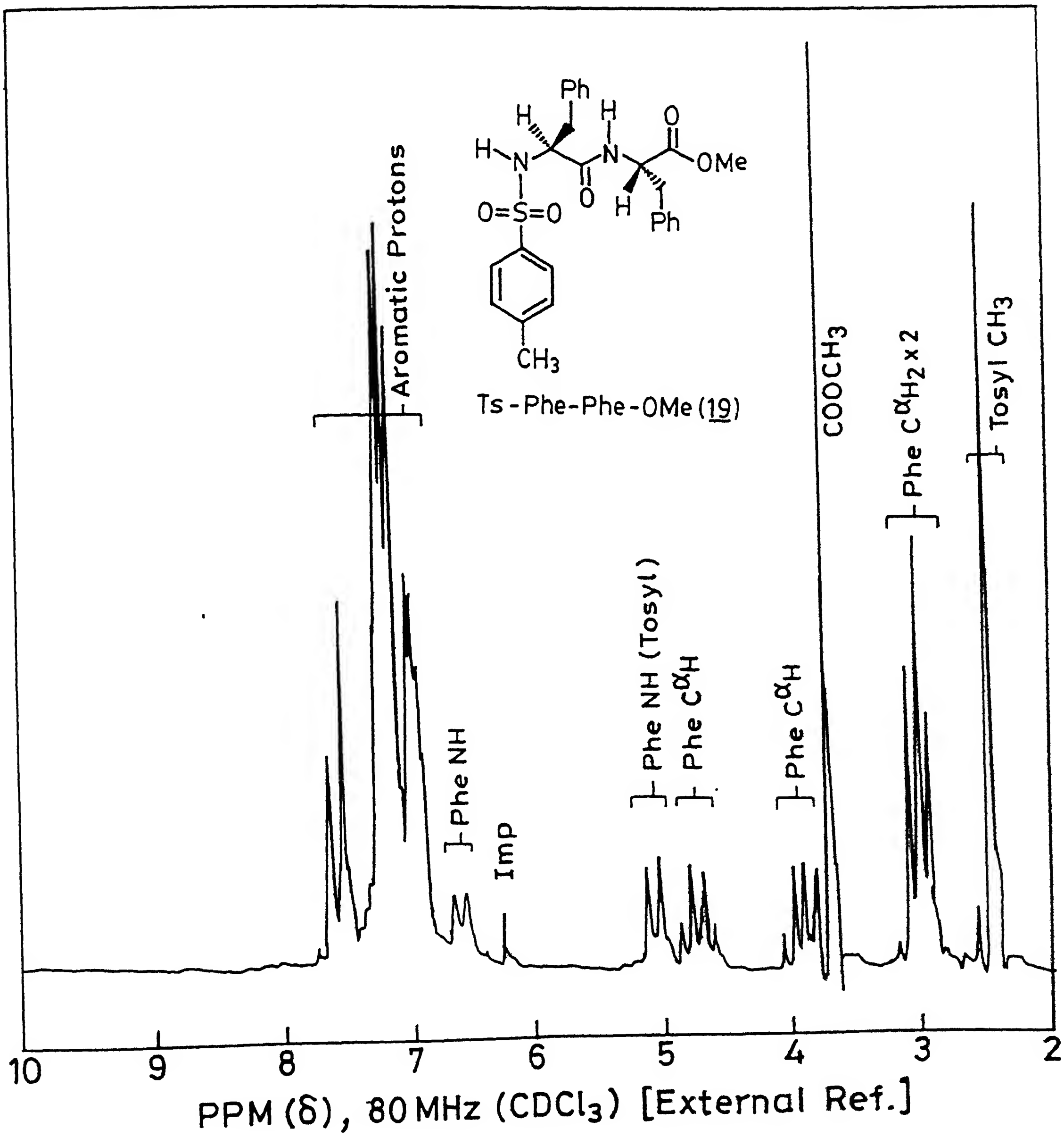

Ts-Leu-Pro-OMe (14)
Aromatic Protons
Leu NH (Tosyl)
Pro CαH + Leu CαH
COOCH3
Pro CδH2
Tosyl CH3
Leu CβH2 + CγH + Pro CβH2 + CγH2
Leu CH3 x 2
9 8 7 6 5 4 3 2 1 0
PPM (δ), 80 MHz (CDCl3) [External Ref.]


Ts - Leu - Leu - OMe (16)
Aromatic Protons
Leu NH
NH (Tosyl)
Leu $C^{\alpha}H$
$COOCH_3$ + Leu $C^{\alpha}H$
Tosyl CH_3
Leu $C^{\beta}H_2$ x 2 + $C^{\gamma}H$ x 2
Leu CH_3 x 4
TMS
9 8 7 6 5 4 3 2 1 0
PPM (δ), 80 MHz ($CDCl_3$)

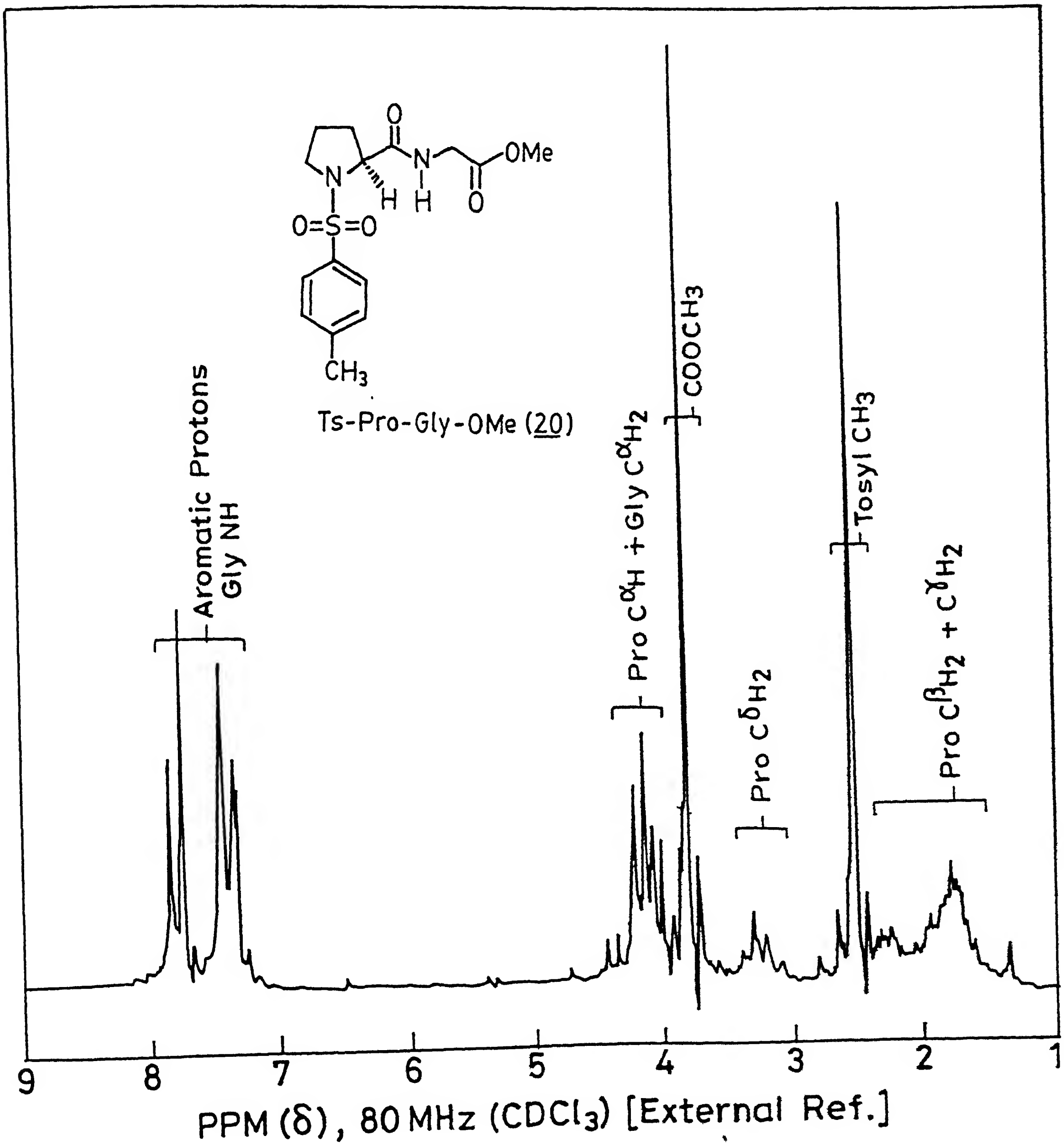

Ph
O
H
OMe
N
H
H
O
O=S=O
CH3
Ts - Pro - Phe - OMe (17)
Aromatic Protons + Phe NH
COOCH3
Pro CδH2 + Phe CβH2
Tosyl CH3
Pro CβH2 + CγH2
Impurity
Pro CαH
Phe CαH
9
8
7
6
5
4
3
2
1
PPM (δ), 80 MHz (CDCl3) [External Ref.]

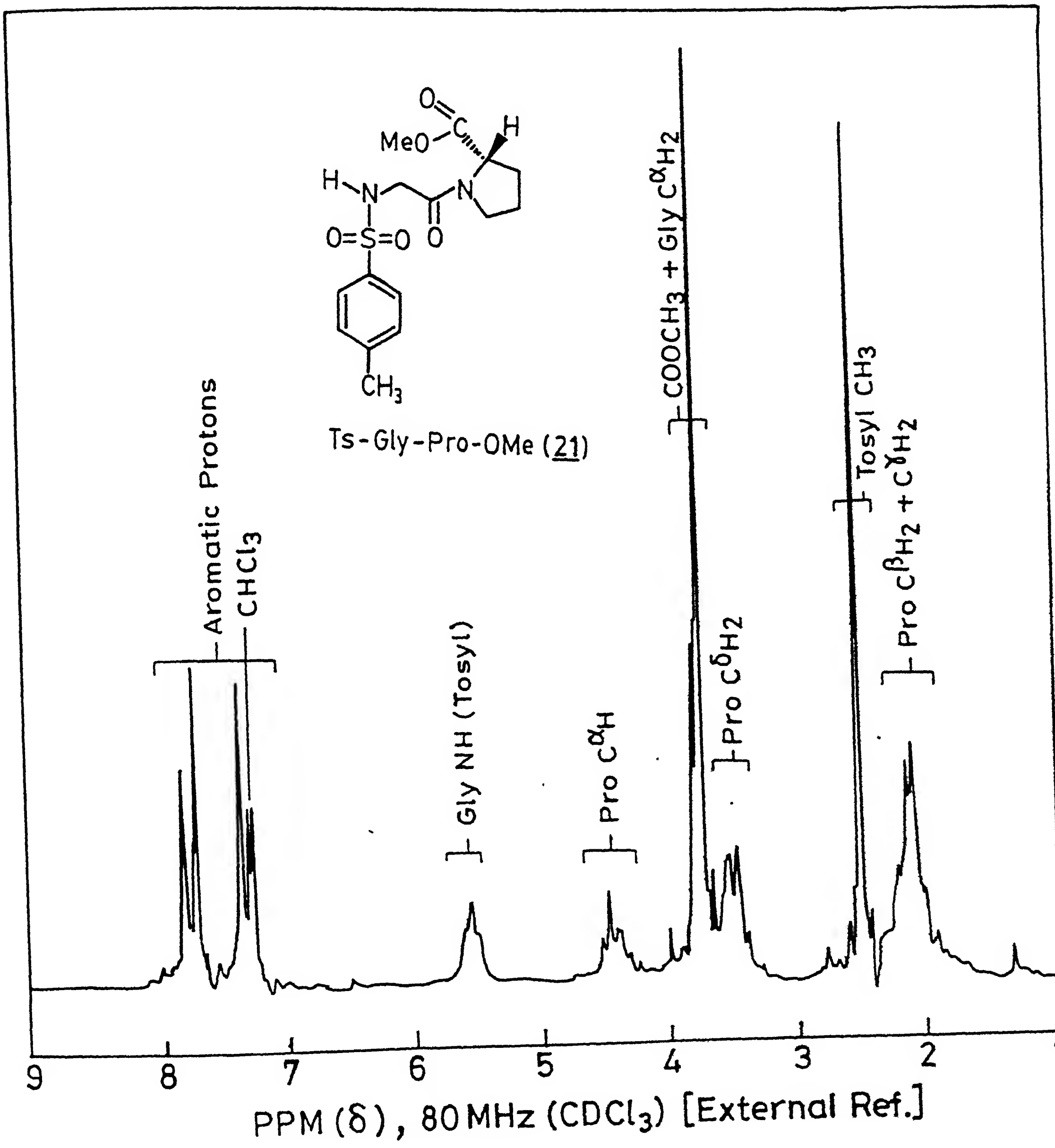


Ts-Phe-Pro-OMe (18)
Aromatic Protons
Phe NH (Tosyl)
Pro CαH + Phe CαH
COOCH3
Pro CδH2 + Phe CβH2
Tosyl CH3
Pro CβH2 + CγH2
9
8
7
6
5
4
3
2
1
PPM (δ), 80 MHz (CDCl3) [External Ref.]

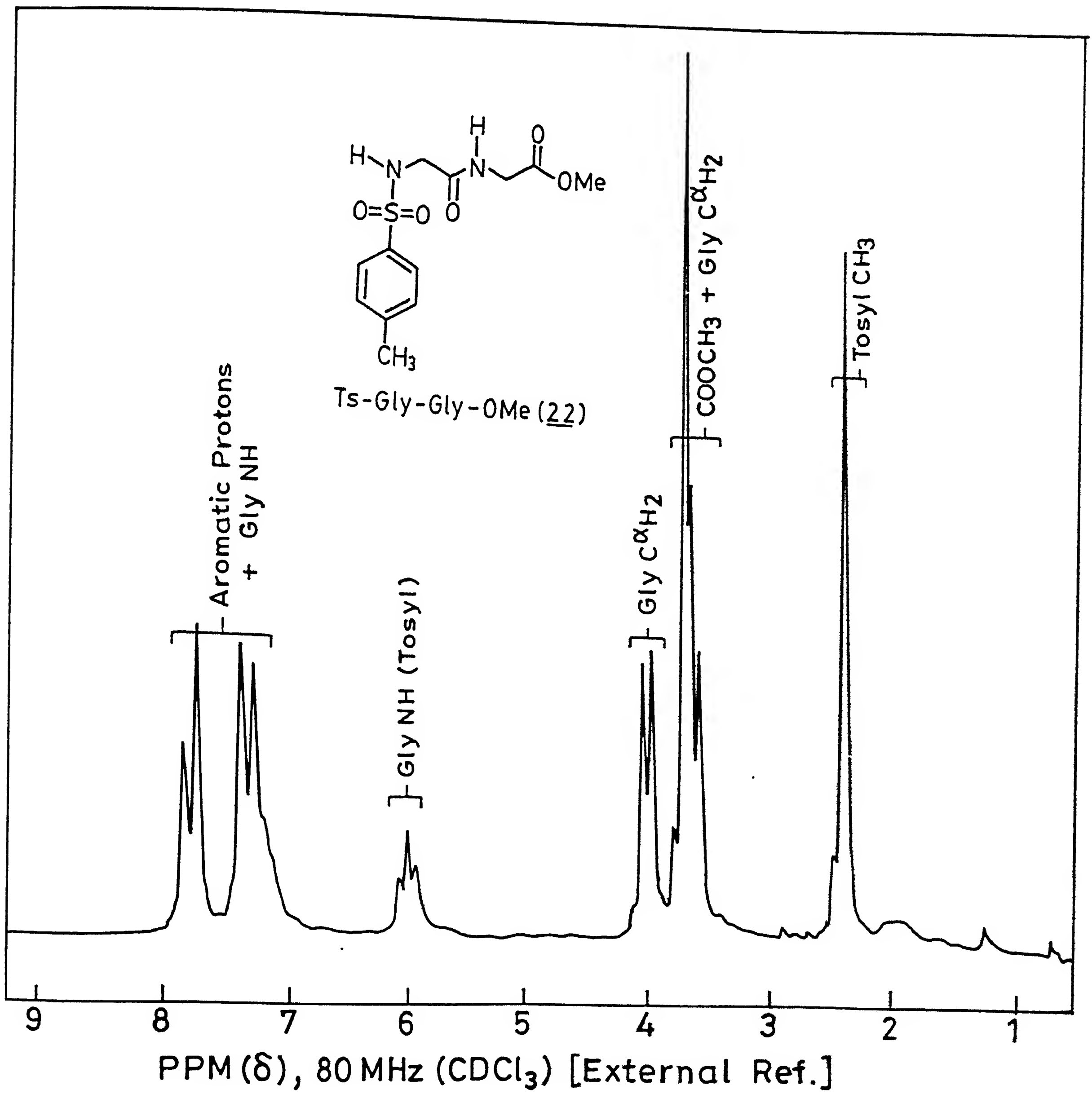


Ts-Phe-Phe-OMe (19)
Aromatic Protons
Phe NH
Imp
Phe NH (Tosyl)
Phe C$^{\alpha}$H
Phe C$^{\alpha}$H
COOCH$_3$
Phe C$^{\alpha}$H$_2$ x 2
Tosyl CH$_3$
10
9
8
7
6
5
4
3
2
PPM (δ), 80 MHz (CDCl$_3$) [External Ref.]

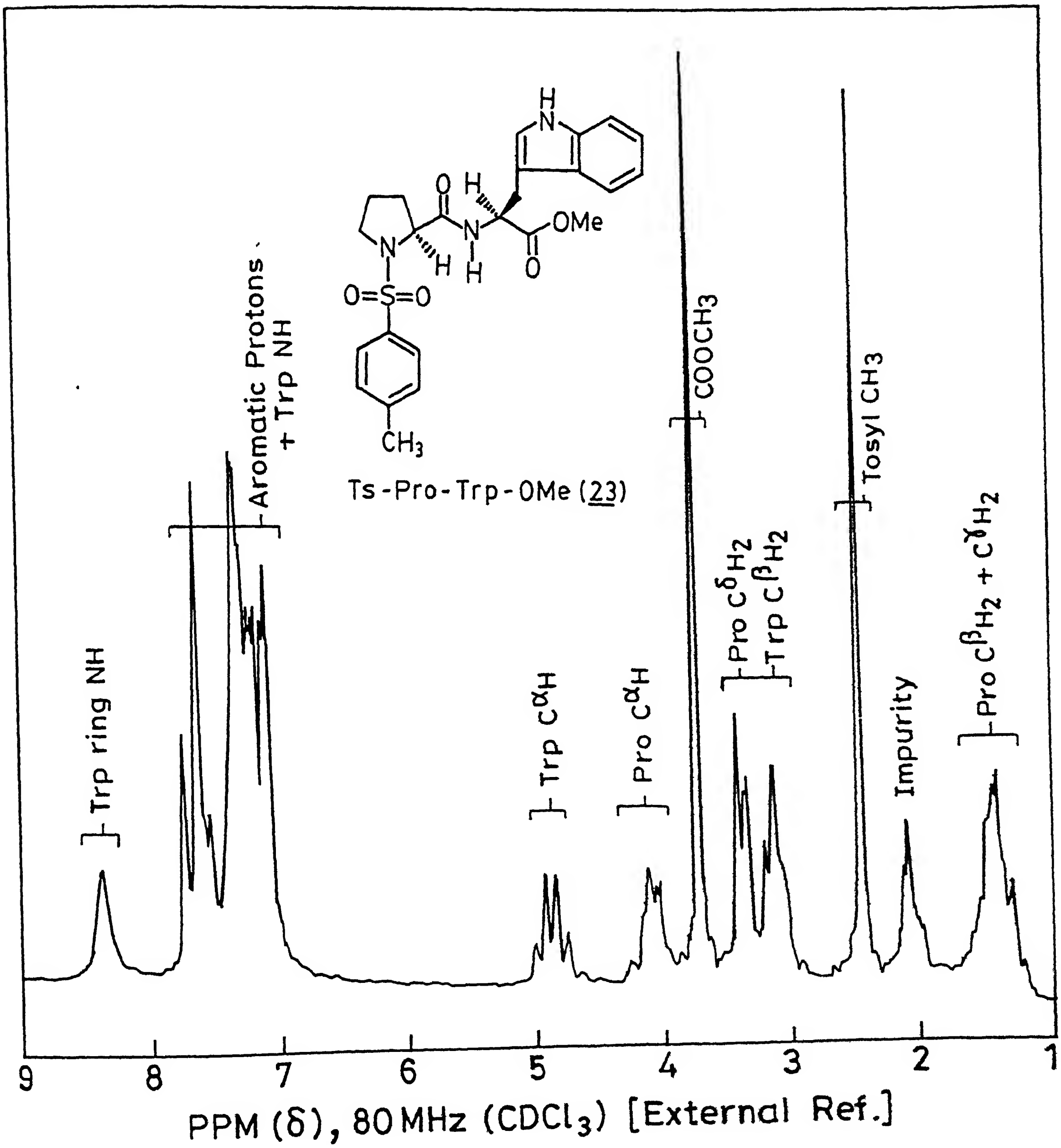

OMe
O=S=O
CH3
Ts-Pro-Gly-OMe (20)
Aromatic Protons
Gly NH
Pro CαH + Gly CαH2
COOCH3
Pro CδH2
Tosyl CH3
Pro CβH2 + CγH2
9
8
7
6
5
4
3
2
1
PPM (δ), 80 MHz (CDCl3) [External Ref.]



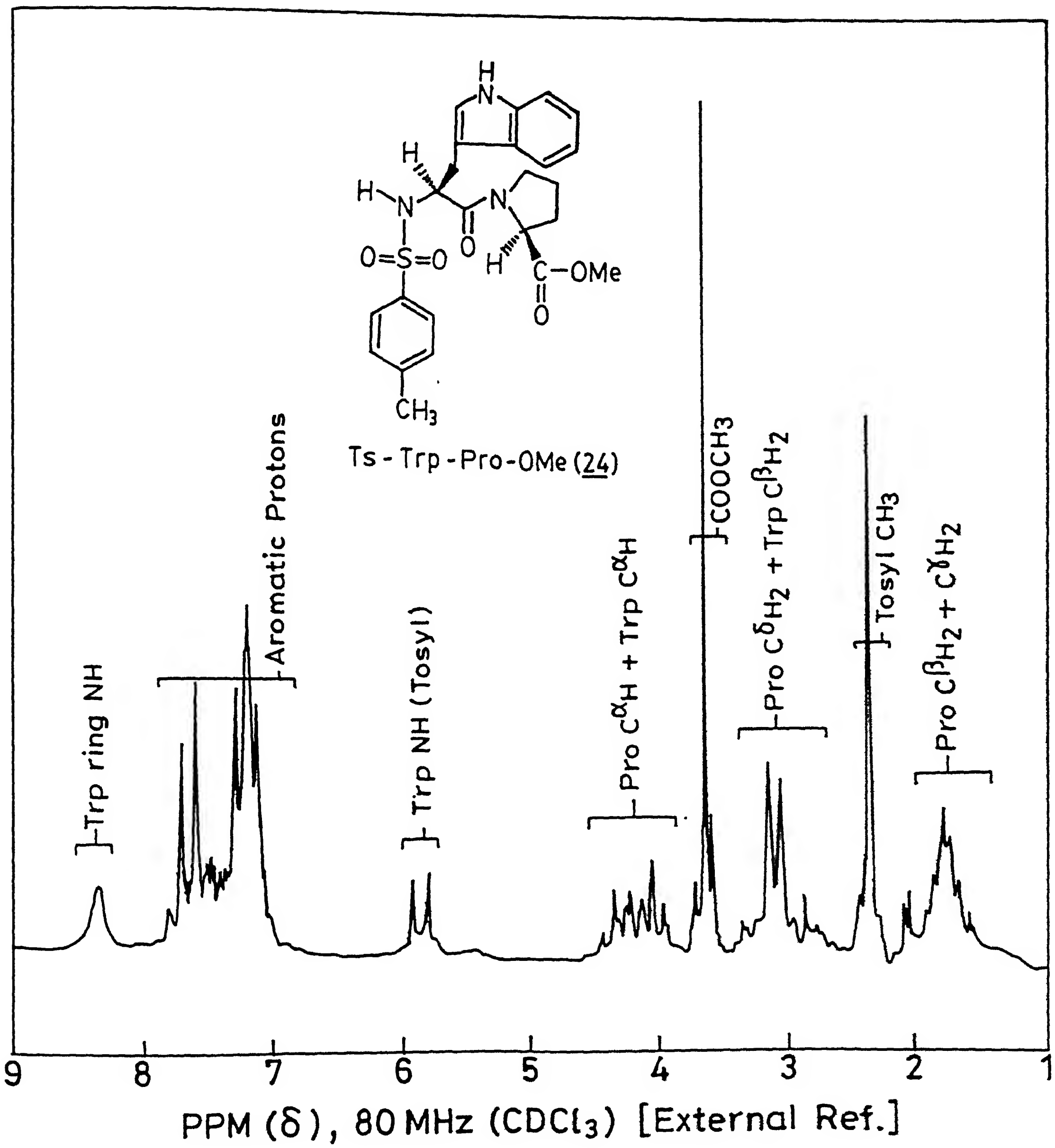
Ts-Gly-Pro-OMe (21)
Aromatic Protons
CHCl3
Gly NH (Tosyl)
Pro CαH
COOCH3 + Gly CαH2
Pro CδH2
Tosyl CH3
Pro CβH2 + CγH2
9
8
7
6
5
4
3
2
1
PPM (δ), 80 MHz (CDCl3) [External Ref.]

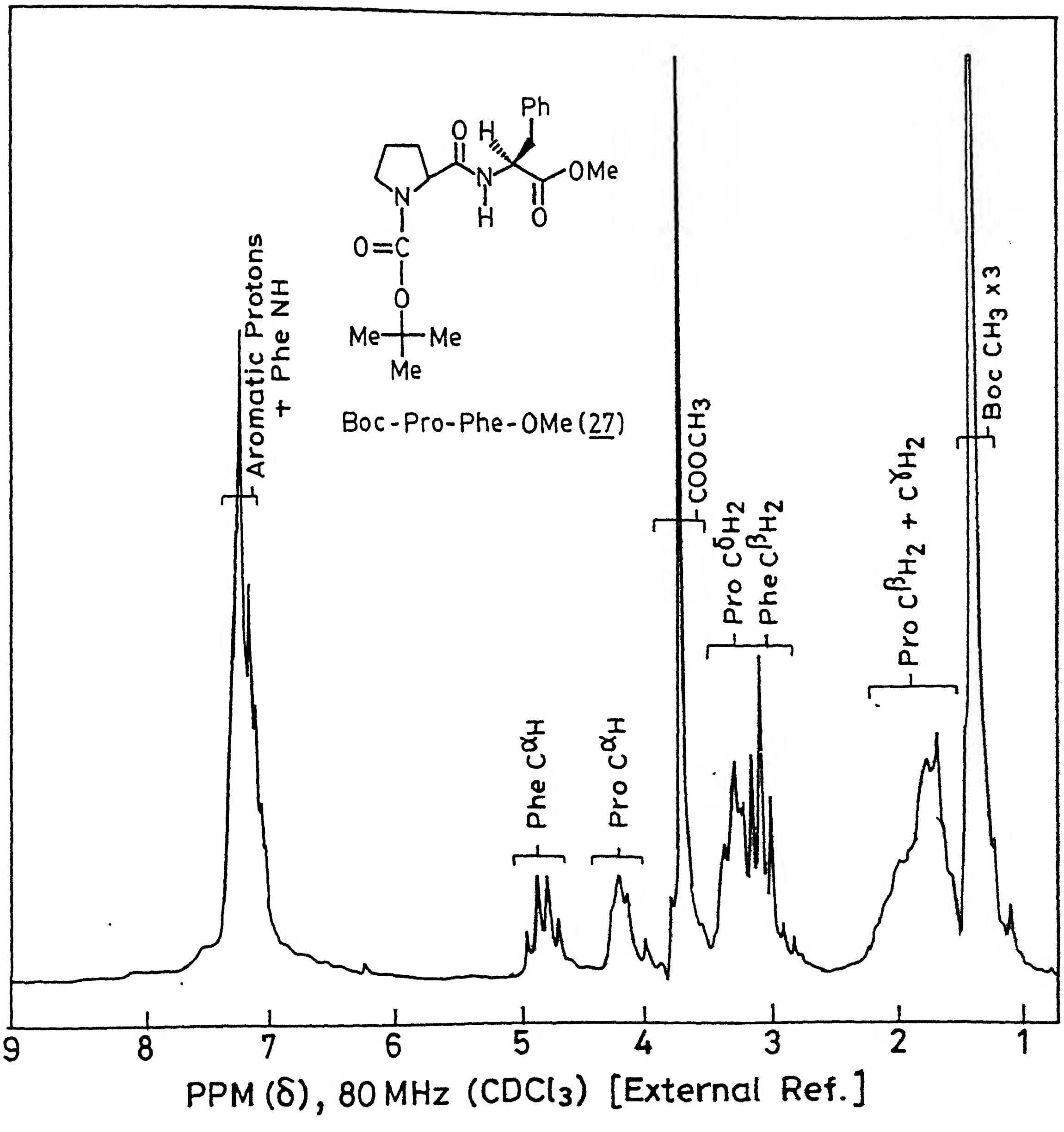


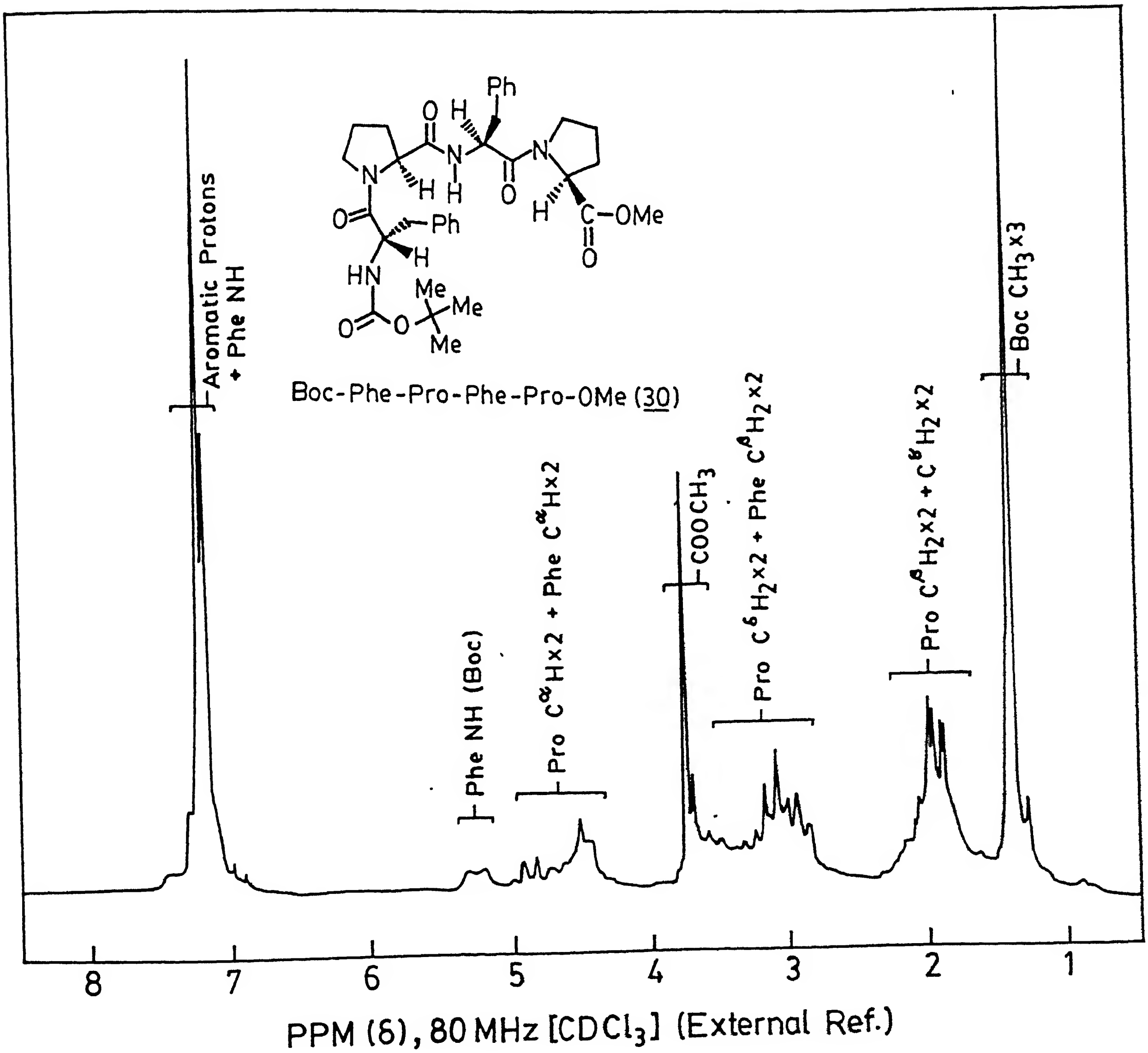
OMe
Ts-Gly-Gly-OMe (22)
Aromatic Protons + Gly NH
Gly NH (Tosyl)
Gly CαH2
COOCH3 + Gly CαH2
Tosyl CH3
9
8
7
6
5
4
3
2
1
PPM (δ), 80 MHz (CDCl3) [External Ref.]

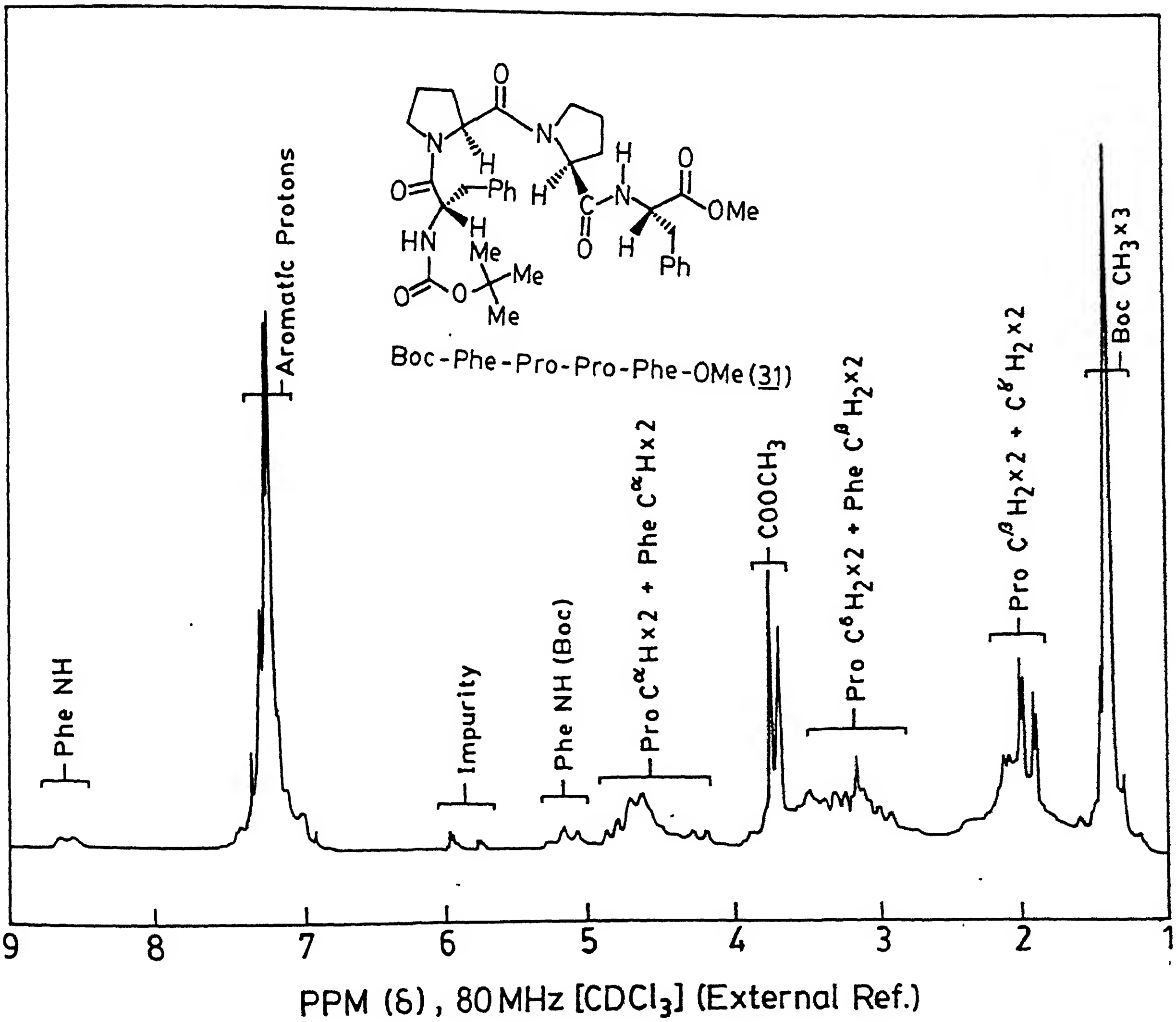


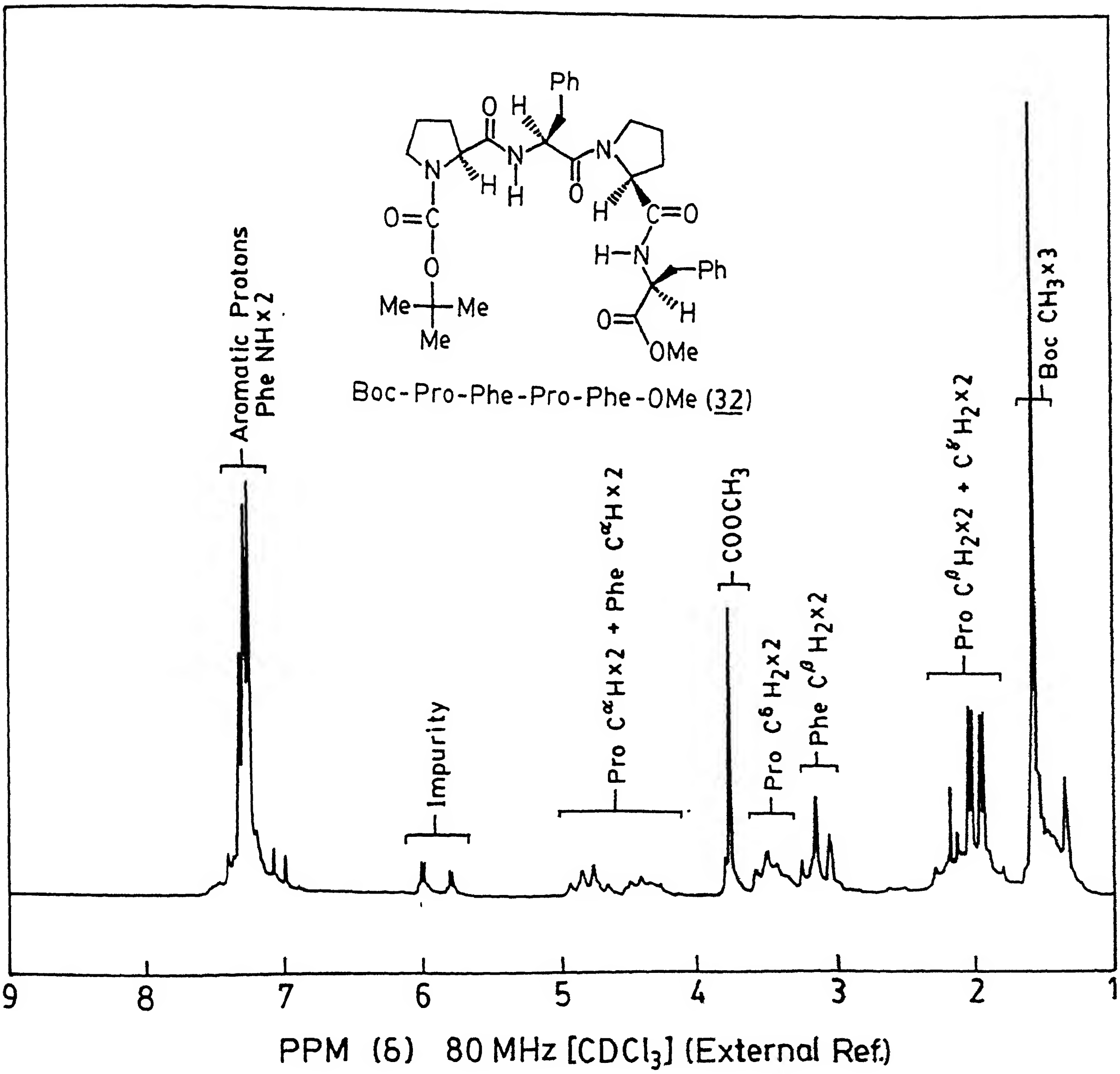
Trp ring NH
Aromatic Protons + Trp NH
Ts-Pro-Trp-OMe (23)
Trp CαH
Pro CαH
COOCH3
Pro CδH2
Trp CβH2
Tosyl CH3
Impurity
Pro CβH2 + CγH2
9
8
7
6
5
4
3
2
1
PPM (δ), 80 MHz (CDCl3) [External Ref.]

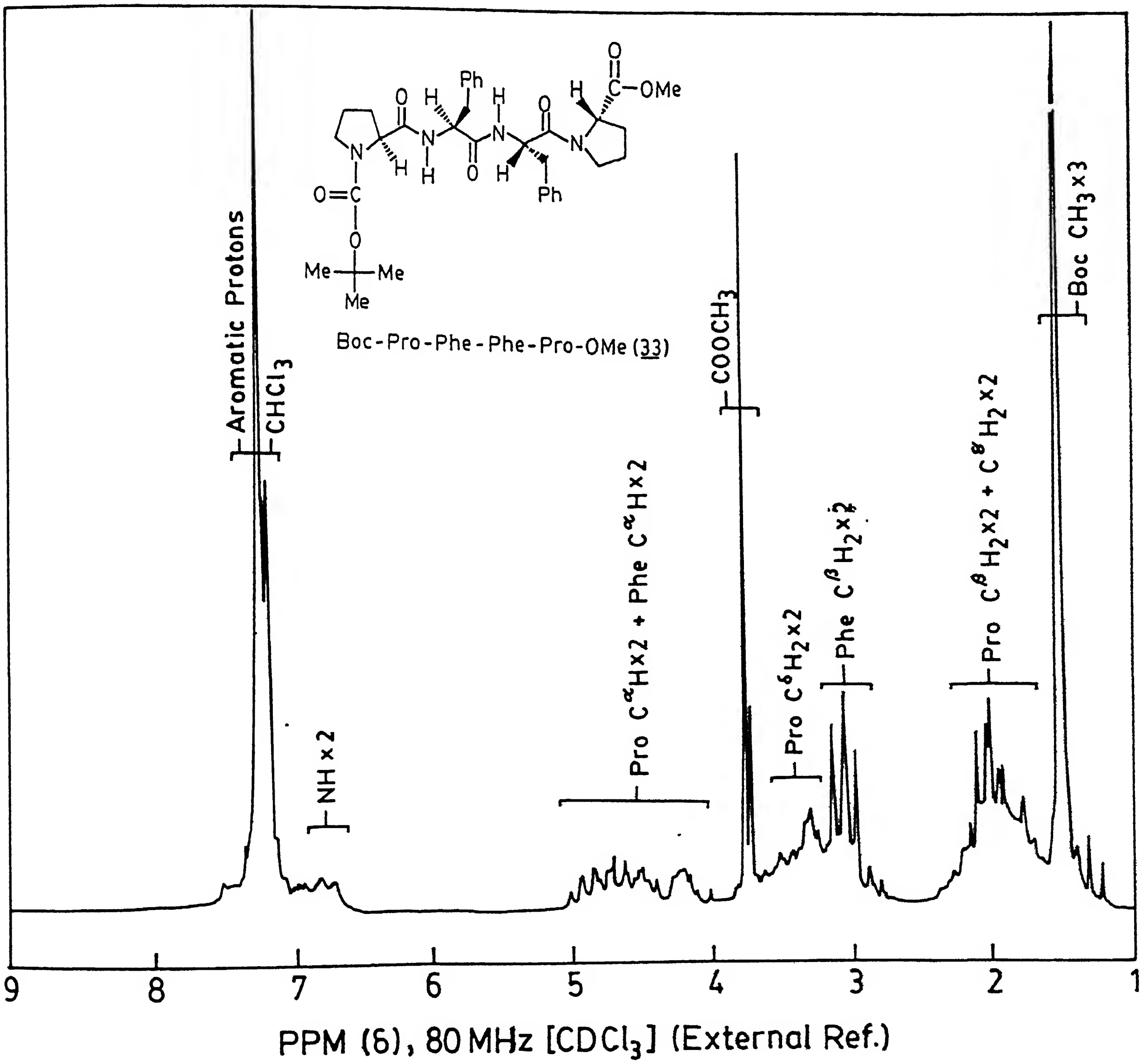

Ts - Trp - Pro - OMe (24)
Trp ring NH
Aromatic Protons
Trp NH (Tosyl)
Pro $C^{\alpha}H$ + Trp $C^{\alpha}H$
$COOCH_3$
Pro $C^{\delta}H_2$ + Trp $C^{\beta}H_2$
Tosyl CH_3
Pro $C^{\beta}H_2$ + $C^{\gamma}H_2$
9
8
7
6
5
4
3
2
1
PPM (δ), 80 MHz ($CDCl_3$) [External Ref.]


Boc-Pro-Phe-OMe (27)
Aromatic Protons + Phe NH
Phe $C^{\alpha}H$
Pro $C^{\alpha}H$
$COOCH_3$
Pro $C^{\delta}H_2$
Phe $C^{\beta}H_2$
Pro $C^{\beta}H_2$ + $C^{\gamma}H_2$
Boc CH_3 x3
9 8 7 6 5 4 3 2 1
PPM (δ), 80 MHz ($CDCl_3$) [External Ref.]


Aromatic Protons + Phe NH
Boc-Phe-Pro-Phe-Pro-OMe (30)
Phe NH (Boc)
Pro $C^{\alpha}H \times 2$ + Phe $C^{\alpha}H \times 2$
$COOCH_3$
Pro $C^{\delta}H_2 \times 2$ + Phe $C^{\beta}H_2 \times 2$
Pro $C^{\beta}H_2 \times 2$ + $C^{\gamma}H_2 \times 2$
Boc $CH_3 \times 3$
8
7
6
5
4
3
2
1
PPM (δ), 80 MHz [$CDCl_3$] (External Ref.)


Boc-Phe-Pro-Pro-Phe-OMe (31)
Phe NH
Aromatic Protons
Impurity
Phe NH (Boc)
Pro $C^{\alpha}H \times 2$ + Phe $C^{\alpha}H \times 2$
$COOCH_3$
Pro $C^{\delta}H_2 \times 2$ + Phe $C^{\beta}H_2 \times 2$
Pro $C^{\beta}H_2 \times 2$ + $C^{\gamma}H_2 \times 2$
Boc $CH_3 \times 3$
9 8 7 6 5 4 3 2 1
PPM (δ), 80 MHz [$CDCl_3$] (External Ref.)


Boc-Pro-Phe-Pro-Phe-OMe (32)
Aromatic Protons
Phe NH x 2
Impurity
Pro CαH x 2 + Phe CαH x 2
COOCH3
Pro CδH2x2
Phe CβH2x2
Pro CβH2x2 + CγH2x2
Boc CH3x3
9
8
7
6
5
4
3
2
1
PPM (δ) 80 MHz [CDCl3] (External Ref.)



Boc-Pro-Phe-Phe-Pro-OMe (33)
Aromatic Protons
CHCl3
NH x 2
Pro CαH x 2 + Phe CαH x 2
COOCH3
Pro CδH2 x 2
Phe CβH2 x 2
Pro CβH2 x 2 + CγH2 x 2
Boc CH3 x 3
9
8
7
6
5
4
3
2
1
PPM (δ), 80 MHz [CDCl3] (External Ref.)

E. EXPERIMENTAL

General:

All amino acids used were of L-configuration. ^1H NMR spectra were obtained on WP 80 Bruker instrument at 80 MHz in $CDCl_3$. The chemical shifts are recorded in ppm with TMS at 0.00 as internal standard or as external reference. IR spectra were recorded on PE 1600 FT instrument either as neat liquid or KBr pellets. FAB mass were recorded using a Jeol SX-120/DA-600 instrument using argon (6KV, 10mA) as the FAB gas. The accelerating voltage was 10KV and the spectra were recorded at room temperature with m-nitrobenzyl alcohol as the matrix. Elemental analysis were carried out in automatic C,H,N analyzer. Silica gel G (Merck) was used for tlc and column chromatography was done on silica gel (acme 100-200 mesh) column, which were generally made from a slurry in benzene or hexane or ethyl acetate. Reactions were monitored whenever possible by tlc. The organic extracts were invariably dried over anhyd. $MgSO_4$ and solvents evaporated *in vacuo*. HPLC was performed on a reverse phase column [Shim-pack CLC-ODS(M)], using UV-VIS spectrophotometric detector [Shimadzu SPD-6AV] at 210 nm.

was poor in CH_2Cl_2. After a period of ~0.25 h, the reaction mixture was admixed with the amino acid methyl ester, freshly prepared at 0^0C from the corresponding ester hydrochloride (1.2 mmol) and triethyl amine (1.2 mmol) in dry CH_2Cl_2 or in a mixture of dry DMF and CH_2Cl_2 . The combined mixture was left stirred at room temperature for 48 h, the precipitated DCU filtered, residue washed with CH_2Cl_2 (2 x 20 mL) and the combined filtrates washed sequentially with cold 2N H_2SO_4 (20 mL), water (20 mL) and saturated bicarbonate solution (20 mL). The organic extract was dried (anhydrous $MgSO_4$) and evaporated *in vacuo.* The residue was, in most cases, directly crystallized from ethylacetate- hexane or purified on a short column of silica gel using ethylacetate-benzene or ethylacetate-hexane as eluents.

C. Deprotection of N, C-Protected Peptides:

(a) N-Deprotection:

(i). (General Procedure-IV): Deprotection of tert-butyloxycarbonyl group:

N-protected peptide methyl ester (1 mmol) was dissolved/suspended in dry CH_2Cl_2 (3 mL), cooled at 0^0C, admixed with CF_3COOH (1 mL), stirred at 0^0C until the starting material disappeared (tlc, ~1.5 h). The solvents were removed under reduced pressure without heating and the residue was thoroughly dried under *vacuum.* The residual trifluoroacetate salt was directly used for the next reaction.

(b) C-Deprotection:

(ii). (General Procedure-V): Hydrolysis of Methyl Ester:

A solution of N-protected peptide methyl ester (1 mmol) in MeOH (~4 mL) was cooled to 0^0C, treated with cold 2N NaOH (4 mL) and stirred at room temperature until the starting material disappeared (tlc, ~1 h). The reaction mixture was concentrated to half the volume without heating *in vacuo,* cooled in ice and acidified (pH~3) with 2N HCl, saturated with solid NaCl and extracted with ethylacetate (3 x 30 mL), dried ($MgSO_4$) and evaporated *in vacuo.* The residue was directly used for the next reaction.

I. Glycine Methyl Ester Hydrochloride (Gly–OMe.HCl)(1):

Dry HCl was passed through a solution of Glycine (10 g, 130 mmol) in absolute (dry) methanol (150 mL) for 0.5 h, concentrated to 20-30 mL and crystallized from dry methanol-ether to give 11.4 g (68%) of Gly-OMe.HCl, mp., 174°C (lit.[40] mp., 175°C).

ir: ν_{max} (KBr)cm^{-1}: 2980 (br, salt), 1740 (ester)

II. L-Leucine Methyl Ester Hydrochloride (L-Leu–OMe.HCl)(2):

Dry HCl was passed through a stirred suspension of L-Leucine (5 g, 38 mmol) in dry methanol (30 mL) for 2 h. The resulting clear solution was evaporated *in vacuo*, the residue dissolved in minimum amount of dry methanol, filtered and dried *in vacuo* to give 5.0 g (71%) of L-Leu-OMe.HCl, mp., 147°C (lit.[41] mp., 151°C).

ir: ν_{max} (KBr)cm^{-1} : 3440 (br, salt), 1730 (ester)

III. L-Phenylalanine Methyl Ester Hydrochloride (L-Phe-OMe.HCl)(3):

Thionyl chloride (4.9 mL, 67 mmol), in drops, followed by L-Phenylalanine (9 g, 54.5 mmol) was added to stirred and ice-cooled dry methanol (45 mL). The mixture was allowed to attain room temperature, refluxed for 2 h, the clear solution evaporated and the residue on crystallization from dry methanol-dry ether gave 10.5 g (89%) of L-Phe-OMe.HCl as white needles, mp., 161°C (lit.[42] mp., 160°C).

ir: ν_{max} (KBr) cm^{-1} : 3320 (salt), 1750 (ester)

IV. L-Proline Methyl Ester Hydrochloride (L-Pro-OMe.HCl)(4):

Dry HCl was passed through a stirred solution of L-Proline (5.75 g, 50 mmol) in dry methanol (60 mL) for 6 h. The resulting clear solution was evaporated *in vacuo*, the residue allowed to stand in a desiccator over P_2O_5 and NaOH overnight to give 5.18 g (63%) of L-Pro-OMe.HCl as thick syrup (lit.[43]).

ir: ν_{max} (neat) cm^{-1}: 2830 (br, salt), 1735 (ester)

V. L-Tryptophan Methyl Ester Hydrochloride (L-Trp-OMe.HCl)(5):

To an ice salt cooled (-10°C) and stirred dry methanol (25 mL) was added, in drops,

$SOCl_2$ (1.9 mL, 23.6 mmol) followed by, rapidly, L-Tryptophan (3.675 g, 18 mmol). From the resulting clear solution, after a short while, a solid precipitated. The reaction mixture was left stirred for an additional 4 h at -5 to 0°C, allowed to attain room temperature, when the precipitated solid redissolveed. The resulting yellow solution was left stirred at room temperature overnight, concentrated *in vacuo* to 5 mL, admixed with dry ether, mixture refrigerated for 4 h, filtered, washed with dry ether and dried to give Trp-OMe.HCl (3.995 g, 84.8%) mp., 214°C (lit.[44] mp., 213°C)

ir : ν_{max} (KBr) cm^{-1} : 3270 (salt), 1740 (ester)

VI. N-Tosyl Glycine (N-Ts-Gly) (6):

Prepared by General Procedure-II

Yield : 69%

mp : 148-150°C (lit.[45] mp., 149-150°C)

ir : ν_{max} (KBr) cm^{-1}: 3271, 3063, 2976, 2927, 1720, 1596, 1497, 1156

VII. N-Tosyl Proline (N-Ts-Pro) (7):

Prepared by General Procedure-II

Yield : 67%

mp : 42-44°C (lit.[46] mp., 39-42°C)

nmr : δ ($CDCl_3$) : 1.59-2.25 (m, 4H, Pro $C^{\beta}H_2$ + $C^{\gamma}H_2$), 2.40 (s, 3H, tosyl CH_3), 3.09-3.68 (m, 2H, Pro $C^{\delta}H_2$), 4.25 (m, 1H, Pro $C^{\alpha}H$), 7.31, 7.71 (d, d, 4H, aromatic protons), 9.43 (s, 1H, COOH).

VIII. N-Tosyl Leucine (N-Ts-Leu) (8):

Prepared by General Procedure-II

Yield : 70.5%

mp : 123°C (lit.[47] mp., 124°C)

nmr : δ ($CDCl_3$) : 0.87 (m, 6H, Leu CH_3 x 2) 1.56 (m, 3H, Leu $C^{\beta}H_2$ + $C^{\gamma}H$), 2.44 (s, 3H, tosyl CH_3), 3.94 (t, 1H, Leu $C^{\alpha}H$), 7.19, 7.94 (d, d, 4H, aromatic protons).

IX. N-Tosyl Phenylalanine (N-Ts-Phe) (9):

Prepared by General Procedure-II

Yield : 72%

mp : 163^{0}C (lit.[48] mp., 164-165^{0}C)

ir : ν_{max} (KBr) cm^{-1}: 3284, 3032, 2924, 1729, 1675, 1596, 1159

X. N-Tosyl Tryptophan (N-Ts-Trp) (10):

Prepared by General Procedure-II

Yield : 62%

mp : 174^{0}C (lit.[49] mp., 176^{0}C)

ir : ν_{max} (KBr) cm^{-1}: 3384, 3296, 3047, 2921, 2854, 1753, 1618, 1598, 1160

XI. N-^{t}Butyloxycarbonyl Proline (N-Boc-Pro) (11):

Prepared by General Procedure-I

Yield : 63%

mp : 135^{0}C (lit.[50] mp., 136-137^{0}C)

ir : ν_{max} (KBr) cm^{-1}: 2973, 2932, 1739, 1637, 1548

XII. N-^{t}Butyloxycarbonyl Phenylalanine (N-Boc-Phe) (12):

Prepared by General Procedure-I

Yield : 56%

mp : 78-80^{0}C (lit[50]. mp., 79-80^{0}C)

ir : ν_{max} (KBr) cm^{-1}: 3384, 2978, 2928, 1706, 1510

XIII. Ts-Pro-Pro-OMe (13):

Obtained by coupling Ts-Pro (7) and Pro-OMe.HCl (4) by DCC/HOBt method (General Procedure-III).

Yield : 56.9%

mp : 105^{0}C

ir : ν_{max} (KBr) cm^{-1} : 2957, 2924, 1748, 1664, 1597, 1150

nmr : δ ($CDCl_3$) : 1.59-2.28 (m, 8H, Pro $C^{\beta}H_2$ x 2 + Pro $C^{\gamma}H_2$ x 2), 2.37 (s, 3H, tosyl CH_3), 3.37 (m, 2H, Pro $C^{\delta}H_2$), 3.69 (s+m, 5H, $COOCH_3$ + Pro $C^{\delta}H_2$), 4.31-4.78 (m, 2H, Pro $C^{\alpha}H$ x 2), 7.28, 7.81 (d,d, 4H, aromatic protons).

ms (FAB) : m/z : 381 $(M+H)^+$

Anal. Calcd. for $C_{18}H_{24}SN_2O_5$: C, 56.84; H, 6.31; N, 7.36 %

Found : C, 57.32; H, 6.42; N, 7.62 %

XIV. Ts-Leu-Pro-OMe (14):

Obtained by coupling Ts-Leu (8) and Pro-OMe.HCl (4) by DCC/HOBt method (General Procedure-III).

Yield : 47.9%

mp : 124-126^{0}C

ir : ν_{max} (KBr) cm^{-1} : 3137, 2953, 1758, 1639, 1598, 1167

nmr : δ ($CDCl_3$) : 0.94 (d, 6H, Leu CH_3 x 2), 1.18-2.15 (brm, 7H, Leu $C^{\beta}H_2$ + $C^{\gamma}H$ + Pro $C^{\beta}H_2$ + Pro $C^{\gamma}H_2$), 2.41 (s, 3H, tosyl CH_3), 3.37 (m, 2H, $C^{\delta}H_2$), 3.65 (s, 3H, $COOCH_3$), 3.72-4.12 (m, 2H, Pro $C^{\alpha}H$ + Leu $C^{\alpha}H$), 5.47 (d, 1H, Leu NH), 7.28, 7.72 (d,d, 4H, aromatic protons).

ms (FAB) : m/z : 397 $(M+H)^+$

XV. Ts-Pro-Leu-OMe (15):

Obtained by coupling Ts-Pro (7) and Leu-OMe.HCl (2) by DCC/HOBt method (General Procedure-III).

Yield : 53.9%

mp : 100-102^{0}C

ir : ν_{max} (KBr) cm^{-1} : 3263, 2956, 1746, 1653, 1559, 1157

nmr : δ ($CDCl_3$) : 0.94 (d, 6H, Leu CH_3 x 2), 1.34-2.0 (m, 7H, Leu $C^{\beta}H_2$ + $C^{\gamma}H$ + Pro $C^{\beta}H_2$ + Pro $C^{\gamma}H_2$), 2.43 (s, 3H, tosyl CH_3), 3.75 (s+m, 5H, $C^{\delta}H_2$ + $COOCH_3$), 4.06 (m, 1H, Pro $C^{\alpha}H$), 4.53 (m, 1H, Leu $C^{\alpha}H$), 7.37, 7.78 (d+m, d, 5H, aromatic protons + Leu NH).

ms (FAB) : m/z : 397 $(M+H)^+$

Anal. Calcd. for $C_{19}H_{28}SN_2O_5$: C, 57.57; H, 7.07; N, 7.07 %

Found : C, 57.50; H, 6.79; N, 6.85 %

XVI. Ts-Leu-Leu-OMe (16):

Obtained by coupling Ts-Leu (8) and Leu-OMe.HCl (2) by DCC/HOBt method (General Procedure-III).

Yield : 59.4%

mp : 124°C (lit.[51] mp., 123-124°C)

ir : ν_{max} (KBr) cm^{-1} : 3261, 2957, 1726, 1657, 1597, 1527, 1166

nmr : δ ($CDCl_3$) : 0.81 (d, 12H, Leu CH_3 x 4), 1.19-1.87 (m, 6H, Leu $C^\beta H_2$ x 2 + $C^\gamma H$ x 2), 2.25 (s, 3H, tosyl CH_3), 3.72 (s+m, 4H, $COOCH_3$ + Leu $C^\alpha H$), 4.37 (m, 1H, Leu $C^\alpha H$), 6.03 (d, 1H, Leu NH), 6.75 (d, 1H, Leu NH), 7.25, 7.78 (d,d, 4H, aromatic protons).

ms (FAB) : m/z : 413 $(M+H)^+$

XVII. Ts-Pro-Phe-OMe (17):

Obtained by coupling Ts-Pro (7) and Phe-OMe.HCl (3) by DCC/HOBt method (General Procedure-III).

Yield : 56.9%

mp : gummy

ir : ν_{max} (neat) cm^{-1} : 3396, 2953, 1744, 1674, 1597, 1517, 1161

nmr : δ ($CDCl_3$) : 1.31-1.66 (m, 4H, Pro $C^\beta H_2$ + Pro $C^\gamma H_2$), 2.41 (s, 3H, tosyl CH_3), 2.88-3.47 (m, 4H, Phe $C^\beta H_2$ + Pro $C^\delta H_2$), 3.75 (s, 3H, $COOCH_3$), 4.03 (m, 1H, Pro $C^\alpha H$), 4.81 (m, 1H, Phe $C^\alpha H$), 7.0-7.81(m+d, 10H, aromatic protons + Phe NH).

ms (FAB) : m/z : 431 $(M+H)^+$

Anal. Calcd. for $C_{22}H_{26}SN_2O_5$: C, 61.39; H, 6.04; N, 6.51 %

Found : C, 61.07; H, 6.03; N, 6.18 %

XVIII. Ts-Phe-Pro-OMe (18):

Obtained by coupling Ts-Phe (9) and Pro-OMe.HCl (4) by DCC/HOBt method (General Procedure-III).

Yield : 50.8%

mp : gummy

ir : ν_{max} (neat) cm^{-1} : 3190, 2924, 1744, 1638, 1598, 1161

nmr : δ ($CDCl_3$) : 1.41-1.87 (m, 4H, Pro $C^{\beta}H_2$ + Pro $C^{\gamma}H_2$), 2.28 (s, 3H, tosyl CH_3), 2.66-3.19 (m, 4H, Phe $C^{\beta}H_2$ + Pro $C^{\delta}H_2$), 3.59 (s, 3H, $COOCH_3$), 3.62-4.25 (m, 2H, Pro $C^{\alpha}H$ + Phe $C^{\alpha}H$), 5.53 (d, 1H, Phe NH), 7.0-7.78 (m+d, 9H, aromatic protons).

ms (FAB) : m/z : 431 $(M+H)^+$

XIX. Ts-Phe-Phe-OMe (19):

Obtained by coupling Ts-Phe (9) and Phe-OMe.HCl (3) by DCC/HOBt method (General Procedure-III).

Yield : 58%

mp : 131-132^{0}C

ir : ν_{max} (KBr) cm^{-1} : 3349, 3317, 3258, 1741, 1662, 1597, 1541, 1159

nmr : δ ($CDCl_3$) : 2.31 (s, 3H, tosyl CH_3), 2.88 (m, 4H, Phe $C^{\beta}H_2$ x 2), 3.59 (s, 3H, $COOCH_3$), 3.78(m, 1H, Phe $C^{\alpha}H$), 4.63 (m, 1H, Phe $C^{\alpha}H$), 4.94 (d, 1H, Phe NH), 6.5 (d, 1H, Phe NH), 6.72-7.59 (m+d, 14H, aromatic protons).

ms (FAB) : m/z : 481 $(M+H)^+$

XX. Ts-Pro-Gly-OMe (20):

Obtained by coupling Ts-Pro (7) and Gly-OMe.HCl (1) by DCC/HOBt method (General Procedure-III).

Yield : 55.8%

mp : gummy

ir : ν_{max} (neat) cm^{-1} : 3390, 2953, 1749, 1672, 1596, 1527, 1160

nmr : δ ($CDCl_3$) : 1.44-2.06 (m, 4H, Pro $C^{\beta}H_2$ + Pro $C^{\gamma}H_2$), 2.53 (s, 3H, tosyl CH_3),

3.0-3.47 (m, 2H, Pro $C^{\delta}H_2$), 3.81 (s, 3H, $COOCH_3$), 3.91-4.44 (m, 3H, Pro $C^{\alpha}H$ +Gly CH_2), 7.37, 7.78 (d+m,d, 5H, aromatic protons + Gly NH).

ms (FAB) : m/z : 341 $(M+H)^+$

XXI. Ts-Gly-Pro-OMe (21):

Obtained by coupling Ts-Gly (6) and Pro-OMe.HCl (4) by DCC/HOBt method (General Procedure-III).

Yield : 50%

mp : 60-61°C

ir : ν_{max} (KBr) cm^{-1} : 3233, 2954, 2926, 1742, 1646, 1160

nmr : δ ($CDCl_3$) : 1.65-2.34 (m, 4H, Pro $C^{\beta}H_2$ + Pro $C^{\gamma}H_2$), 2.47 (s, 3H, tosyl CH_3), 3.34-3.62 (m, 2H, Pro $C^{\delta}H_2$), 3.69 (s+d, 5H, $COOCH_3$ + Gly CH_2), 4.41 (m, 1H, Pro $C^{\alpha}H$), 5.53 (m, 1H, Gly NH), 7.31, 7.78 (d,d, 4H, aromatic protons).

ms (FAB) : m/z : 341 $(M+H)^+$

Anal. Calcd. for $C_{15}H_{20}SN_2O_5$: C, 52.94; H, 5.88; N, 8.23 %

Found : C, 52.69; H, 5.92; N, 7.83 %

XXII. Ts-Gly-Gly-OMe (22):

Obtained by coupling Ts-Gly (6) and Gly-OMe.HCl (1) by DCC/HOBt method (General Procedure-III).

Yield : 54.6%

mp : 91-92°C

ir : ν_{max} (KBr) cm^{-1} : 3412, 3143, 1746, 1646, 1536, 1165

nmr : δ ($CDCl_3$) : 2.41 (s, 3H, tosyl CH_3), 3.65 (s+m, 5H, $COOCH_3$ + Gly CH_2), 4.0 (d, 2H, Gly CH_2), 6.03 (m, 1H, Gly NH), 7.31, 7.78 (d+m,d, 5H, aromatic protons + Gly NH).

ms (FAB) : m/z : 301 $(M+H)^+$

Anal. Calcd. for $C_{12}H_{16}SN_2O_5$: C, 48.0; H, 5.33; N, 9.33 %

Found : C, 48.27; H, 5.33; N, 9.24 %

XXIII. Ts-Pro-Trp-OMe (23):

Obtained by coupling Ts-Pro (7) and Trp-OMe.HCl (5) by DCC/HOBt method (General Procedure-III).

Yield : 53%

mp : 67-69^{0}C

ir : ν_{max} (KBr) cm^{-1} : 3394, 2924, 1742, 1666, 1596, 1520, 1159

nmr : δ ($CDCl_3$) : 1.03-1.63 (m, 4H, Pro $C^\beta H_2$ + Pro $C^\gamma H_2$), 2.31 (s, 3H, tosyl CH_3), 3.0, 3.25 (m,m, 4H, Trp $C^\beta H_2$ + Pro $C^\delta H_2$), 3.63 (s, 3H, $COOCH_3$), 4.0 (m, 1H, Pro $C^\alpha H$), 4.75 (m, 1H, Trp $C^\alpha H$), 6.81-7.69 (m, 10H, aromatic protons + Trp NH), 8.22 (brs, 1H, Trp ring NH).

ms (FAB) : m/z : 470 $(M+H)^+$

XXIV. Ts-Trp-Pro-OMe (24):

Obtained by coupling Ts-Trp (10) and Pro-OMe.HCl (4) by DCC/HOBt method (General Procedure-III).

Yield : 61.9%

mp : 71-73^{0}C

ir : ν_{max} (KBr) cm^{-1} : 3396, 2923, 1742, 1636, 1558, 1160

nmr : δ ($CDCl_3$) : 1.47-1.94 (m, 4H, Pro $C^\beta H_2$ + Pro $C^\gamma H_2$), 2.34 (s, 3H, tosyl CH_3), 2.60-3.37 (d+m, 4H, Trp $C^\beta H_2$ + Pro $C^\delta H_2$), 3.66 (s, 3H, $COOCH_3$), 3.94-4.47 (m, 2H, Pro $C^\alpha H$ + Trp $C^\alpha H$), 5.84 (d, 1H, Trp NH), 7.0-7.81 (brm, 9H, aromatic protons), 8.31 (brs, 1H, Trp ring NH).

ms (FAB) : m/z : 470 $(M+H)^+$

XXV. Ts-Trp-Trp-OMe (25):

Obtained by coupling Ts-Trp (10) and Trp-OMe.HCl (5) by DCC/HOBt method (General Procedure-III).

Yield : 46.9%

mp : 120-121^{0}C

ir : ν_{max} (KBr) cm^{-1} : 3402, 2922, 1750, 1671, 1523, 1162

nmr : δ ($CDCl_3$) : 2.31 (s, 3H tosyl CH_3), 3.12 (m, 4H, Trp $C^{\beta}H_2$ x 2), 3.62 (s, 3H, $COOCH_3$), 3.78-4.18 (m, 1H, Trp $C^{\alpha}H$), 4.59-5.09 (d+m, 2H, Trp $C^{\alpha}H$ + Trp NH), 6.59-8.12 (m, 17H, aromatic protons + Trp NH + Trp ring NH x 2).

ms (FAB) : m/z : 559 $(M+H)^+$

XXVI. Boc-Phe-Pro-OMe (<u>26</u>):

Obtained by coupling Boc-Phe (<u>12</u>) and Pro-OMe.HCl (<u>4</u>) by DCC/HOBt method (General Procedure-III).

Yield : 59%

mp : gummy

ir : ν_{max} (neat) cm^{-1} : 3313, 2975, 1745, 1708, 1647, 1499

nmr : δ ($CDCl_3$) : 1.40 (s, 9H, Boc CH_3 x 3), 1.71-2.28 (m, 4H, Pro $C^{\beta}H_2$ + Pro $C^{\gamma}H_2$), 3.0 (m, 2H, Phe $C^{\beta}H_2$), 3.46-3.84 (s+m, 5H, Pro $C^{\delta}H_2$ + $COOCH_3$), 4.34-4.78 (m, 2H, Pro $C^{\alpha}H$ + Phe $C^{\alpha}H$), 5.31 (d, 1H, Phe NH), 7.31 (s, 5H, aromatic protons).

XXVII. Boc-Pro-Phe-OMe (<u>27</u>):

Obtained by coupling Boc-Pro (<u>11</u>) and Phe-OMe.HCl (<u>3</u>) by DCC/HOBt method (General Procedure-III).

Yield : 54%

mp : 74-75°C (lit.[52] mp., 74-76°C)

ir : ν_{max} (KBr) cm^{-1} : 3384, 3080, 2973, 1749, 1694, 1657, 1554

nmr : δ ($CDCl_3$) : 1.44 (s, 9H, Boc CH_3 x 3), 1.53-2.25 (m, 4H, Pro $C^{\beta}H_2$ + Pro $C^{\gamma}H_2$), 3.13-3.31 (brm, 4H, Phe $C^{\beta}H_2$ + Pro $C^{\delta}H_2$), 3.72 (s, 3H, $COOCH_3$), 4.25 (m, 1H, Pro $C^{\alpha}H$), 4.84 (m, 1H, Phe $C^{\alpha}H$), 7.22 (m, 6H, aromatic protons + Phe NH).

XXVIII. Boc-Phe-Pro-OH (<u>28</u>):

Obtained by hydrolysis of Boc-Phe-Pro-OMe (<u>26</u>) by (General Procedure-V)

Yield : 95%

mp : gummy

nmr : δ ($CDCl_3$) : 1.25 (s, 9H, Boc CH_3 x 3), 1.56-2.25 (m, 4H, Pro $C^{\beta}H_2$ + Pro $C^{\gamma}H_2$), 2.94 (m, 2H, Phe $C^{\beta}H_2$), 3.47 (m, 2H, Pro $C^{\delta}H_2$), 4.25 (m, 2H, Pro $C^{\alpha}H$ + Phe $C^{\alpha}H$), 5.37 (d, 1H, Boc NH), 7.13 (s, 5H, aromatic protons).

XXIX. Boc-Pro-Phe-OH (29):

Obtained by hydrolysis of Boc-Pro-Phe-OMe (27) by (General Procedure-V)

Yield : 95%

mp : gummy

nmr : δ ($CDCl_3$) : 1.41 (s, 9H, Boc CH_3 x 3), 2.18 (brm, 4H, Pro $C^{\beta}H_2$ + Pro $C^{\gamma}H_2$), 3.28 (m, 4H, Phe $C^{\beta}H_2$ + Pro $C^{\delta}H_2$), 4.25 (m, 1H, Pro $C^{\alpha}H$), 4.87 (m, 1H, Phe $C^{\alpha}H$), 7.31 (m, 6H, aromatic protons + Phe NH).

XXX. Boc-Phe-Pro-Phe-Pro-OMe (30):

Obtained by coupling Boc-Phe-Pro-OH (28) and Phe-Pro-OMe [freshly prepared from Boc-Phe-Pro-OMe (26) by (General Procedure-IV)] by DCC/HOBt method (General Procedure-III)

Yield : 42.8%

mp : 72-75^{0}C

ir : ν_{max} (KBr) cm^{-1} : 3307, 2976, 2879, 1743, 1705, 1639

nmr : δ ($CDCl_3$) : 1.38 (s, 9H, Boc CH_3 x 3), 1.75-2.28 (m, 8H, Pro $C^{\beta}H_2$ x 2 + Pro $C^{\gamma}H_2$ x 2), 2.78-3.41 (m, 8H, Phe $C^{\beta}H_2$ x 2 + Pro $C^{\delta}H_2$ x 2), 3.72 (s, 3H, $COOCH_3$), 4.28-5.0 (m, 4H, Pro $C^{\alpha}H$ x 2 + Phe $C^{\alpha}H$ x 2), 5.25 (d, 1H, Phe NH), 7.25 (s+m, 11H, aromatic protons + Phe NH).

ms (FAB) : m/z : 621 $(M+H)^+$

XXXI. Boc-Phe-Pro-Pro-Phe-OMe (31):

Obtained by coupling Boc-Phe-Pro-OH (28) and Pro-Phe-OMe [freshly prepared from Boc-Pro-Phe-OMe (27) by (General Procedure-IV)] by DCC/HOBt method (Gen-

eral Procedure-III)

Yield : 39.8%

mp : gummy

ir : ν_{max} (neat) cm^{-1} : 3300, 2976, 2878, 1743, 1707, 1640

nmr : δ ($CDCl_3$) : 1.34 (s, 9H, Boc CH_3 x 3), 1.75-2.19 (m, 8H, Pro $C^{\beta}H_2$ x 2 + Pro $C^{\gamma}H_2$ x 2), 2.78-3.50 (m, 8H, Phe $C^{\beta}H_2$ x 2 + Pro $C^{\delta}H_2$ x 2), 3.69 (s, 3H, $COOCH_3$), 4.09-4.81 (m, 4H, Pro $C^{\alpha}H$ x 2 + Phe $C^{\alpha}H$ x 2), 5.06 (m, 1H, Phe NH). 7.22 (m, 10H, aromatic protons), 8.56 (d, 1H, Phe NH).

ms (FAB) : m/z : 621 $(M+H)^+$

XXXII. Boc-Pro-Phe-Pro-Phe-OMe (32):

Obtained by coupling Boc-Pro-Phe-OH (29) and Pro-Phe-OMe [freshly prepared from Boc-Pro-Phe-OMe (27) by (General Procedure-IV)] by DCC/HOBt method (General Procedure-III)

Yield : 46%

mp : gummy

ir : ν_{max} (neat) cm^{-1} : 3294, 2975, 1744, 1679, 1529

nmr : δ ($CDCl_3$) : 1.46 (s, 9H, Boc CH_3 x 3), 1.68-2.22 (m, 8H, Pro $C^{\beta}H_2$ x 2 + Pro $C^{\gamma}H_2$ x 2), 3.09, 3.44 (m, 8H, Phe $C^{\beta}H_2$ x 2 + Pro $C^{\delta}H_2$ x 2), 3.72 (s, 3H, $COOCH_3$), 4.09-4.90 (m, 4H, Pro $C^{\alpha}H$ x 2 + Phe $C^{\alpha}H$ x 2), 7.19 (m, 12H, aromatic protons + Phe NH x 2).

ms (FAB) : m/z : 621 $(M+H)^+$

XXXIII. Boc-Pro-Phe-Phe-Pro-OMe (33):

Obtained by coupling Boc-Pro-Phe-OH (29) and Phe-Pro-OMe [freshly prepared from Boc-Phe-Pro-OMe (26) by (General Procedure-IV)] by DCC/HOBt method (General Procedure-III)

Yield : 48.9%

mp : 77-79^{0}C

ir : ν_{max} (KBr) cm^{-1} : 3299, 3061, 2977, 2881, 1744, 1686, 1640, 1514

nmr : δ ($CDCl_3$) : 1.50 (s, 9H, Boc CH_3 x 3), 1.65-2.25 (m, 8H, Pro $C^{\beta}H_2$ x 2 + Pro $C^{\gamma}H_2$ x 2), 2.72-3.59 (brm, 8H, Phe $C^{\beta}H_2$ x 2 + Pro $C^{\delta}H_2$ x 2), 3.78 (s, 3H, $COOCH_3$), 4.09-5.03 (brm, 4H, Pro $C^{\alpha}H$ x 2 + Phe $C^{\alpha}H$ x 2), 6.78 (d, 2H, Phe NH x 2), 7.21 (s, 10H, aromatic protons).

ms (FAB) : m/z : 621 $(M+H)^+$

(General Procedure-VI) : Reaction of Proline and target amino acid (AA) with 3 equivalent amounts of water soluble carbodiimide in water :

Aqueous solutions of Proline (1 mmol, 5 mL), target amino acid (1 mmol, 5 mL) and the water soluble carbodiimide, 1, Cyclohexyl-3-(2-morpholinoethyl) carbodiimide metho-p-toluenesulfonate (3 mmol, 5 mL) were mixed and stirred for 48 h at room temperature. The reaction mixture was treated with NaOH (0.6 g, 2 mmol), cooled, tosyl chloride (0.39 g, 2 mmol) was added, left stirred for 4 h at room temperature, filtered (to remove unreacted tosyl chloride), aqueous portion cooled in ice, acidified with 5N HCl to pH 2, saturated with NaCl, extracted with ethylacetate (3 x 20 mL), dried over $MgSO_4$, evaporated, the residue dissolved in minimum amount of MeOH, cooled, treated with ethereal CH_2N_2, evaporated. HPLC performed on residue using reverse phase column [Shim-pack CLC-ODS(M)] and dipeptide formation compared with authentic possible dipeptides.

(General Procedure-VII) : Reaction of N, C-protected Proline and N, C-protected target amino acid (AA) with water soluble carbodiimide in water or water-acetonitrile mixture (4:1) :

Ts-Pro (1 mmol), Ts-AA (1 mmol), Pro-OMe (1 mmol) and AA-OMe (1 mmol) were dissolved in 20 mL water or in water (16 mL), acetonitrile (4 mL) mixture and 2

mmol of water soluble carbodiimide, 1, Cyclohexyl-3-(2-morpholinoethyl) carbodiimide metho-p-toluenesulfonate added. The reaction mixture was left stirred for 48 h at room temperature, extracted with methylene chloride (3 x 20 mL) and the combined extracts were successively washed with 1N HCl (2 x 20 mL), 1N $KHCO_3$ (2 x 20 mL) and water (2 x 20 mL). The organic solution was dried over $MgSO_4$ and evaporated to dryness. HPLC performed on the residue using reverse phase column [Shim-pack CLC-ODS (M)] and compared with the authentic possible dipeptides to determine the peptide formed in water.

XXXIV. Reaction of Proline and Leucine with WSCDI in water: (General Procedure-VI) HPLC analysis of the reaction product (Figure.C.I.1b and 2b): Mobile Phase : $MeCN:H_2O$ = 60:40; Flow rate : 0.8 mL/min; Wavelength = 210 nm

Dipeptide	RT	% in the reaction mixture	% Yield	Ratio of dipeptide
Ts-Pro-Pro-OMe	6.28	-	-	-
Ts-Pro-Leu-OMe	10.55	25	11	41
Ts-Leu-Pro-OMe	9.29	37	16	59
Ts-Leu-Leu-OMe	12.58	-	-	-

XXXV.Reaction of Ts-Pro, Ts-Leu, Pro-OMe and Leu-OMe with WSCDI in water : (General Procedure-VII) HPLC analysis of the reaction product (Figure.C.I.1b and 2b): Mobile Phase : $MeCN:H_2O$ = 60:40; Flow rate : 0.8 mL/min; Wavelength = 210 nm

Dipeptide	RT	% in the reaction mixture	% Yield	Ratio of dipeptide
Ts-Pro-Pro-OMe	6.28	-	-	-
Ts-Pro-Leu-OMe	10.55	38	37	54
Ts-Leu-Pro-OMe	9.29	-	-	-
Ts-Leu-Leu-OMe	12.58	34	32	46

XXXVI. Reaction of Ts-Pro, Ts-Leu, Pro-OMe and Leu-OMe with WSCDI in water : acetonitrile (4:1) : (General Procedure-VII) HPLC analysis of the reaction product (Figure.C.I.1b and 2b): Mobile Phase : MeCN:H_2O = 60:40; Flow rate : 0.8 mL/min; Wavelength = 210 nm

Dipeptide	RT	% in the reaction mixture	% Yield	Ratio of dipeptide
Ts-Pro-Pro-OMe	6.28	05	05	05
Ts-Pro-Leu-OMe	10.55	54	56	62
Ts-Leu-Pro-OMe	9.29	-	-	-
Ts-Leu-Leu-OMe	12.58	30	30	33

XXXVII. Reaction of Proline and Phenylalanine with WSCDI in water: (General Procedure-VI) HPLC analysis of the reaction product (Figure.C.I.1c and 2c): Mobile Phase : MeCN:H_2O = 60:40; Flow rate : 0.8 mL/min; Wavelength = 210 nm

Dipeptide	RT	% in the reaction mixture	% Yield	Ratio of dipeptide
Ts-Pro-Pro-OMe	6.26	-	-	-
Ts-Pro-Phe-OMe	11.16	10	4.4	26
Ts-Phe-Pro-OMe	9.25	29	12.7	74
Ts-Phe-Phe-OMe	13.67	-	-	-

XXXVIII. Reaction of Ts-Pro, Ts-Phe, Pro-OMe and Phe-OMe with WSCDI in water : (General Procedure-VII) HPLC analysis of the reaction product (Figure.C.I.1c and 2c): Mobile Phase : MeCN:H_2O = 60:40; Flow rate : 0.8 mL/min; Wavelength = 210 nm

Dipeptide	RT	% in the reaction mixture	% Yield	Ratio of dipeptide
Ts-Pro-Pro-OMe	6.26	-	-	-
Ts-Pro-Phe-OMe	11.16	42	47	53
Ts-Phe-Pro-OMe	9.25	-	-	-
Ts-Phe-Phe-OMe	13.67	42	42	47

XXXIX. Reaction of Ts-Pro, Ts-Phe, Pro-OMe and Phe-OMe with WSCDI in water : acetonitrile (4:1) : (General Procedure-VII) HPLC analysis of the reaction product (Figure.C.I.1c and 2c): Mobile Phase : MeCN:H_2O = 60:40; Flow rate : 0.8 mL/min; Wavelength = 210 nm

Dipeptide	RT	% in the reaction mixture	% Yield	Ratio of dipeptide
Ts-Pro-Pro-OMe	6.26	-	-	-
Ts-Pro-Phe-OMe	11.16	53	68	56
Ts-Phe-Pro-OMe	9.25	-	-	-
Ts-Phe-Phe-OMe	13.67	47	54	44

XXXX. Reaction of Proline and Glycine with WSCDI in water: (General Procedure-VI) HPLC analysis of the reaction product (Figure.C.I.1a and 2a): Mobile Phase : MeOH:H_2O = 60:40; Flow rate : 0.8 mL/min; Wavelength = 210 nm

Dipeptide	RT	% in the reaction mixture	% Yield	Ratio of dipeptide
Ts-Pro-Pro-OMe	8.85	-	-	-
Ts-Pro-Gly-OMe	6.86	21	7.3	100
Ts-Gly-Pro-OMe	6.39	-	-	-
Ts-Gly-Gly-OMe	5.25	-	-	-

XXXXI. Reaction of Ts-Pro, Ts-Gly, Pro-OMe and Gly-OMe with WSCDI in water : (General Procedure-VII) HPLC analysis of the reaction product (Figure.C.I.1a and 2a): Mobile Phase : MeOH:H_2O = 60:40; Flow rate : 0.8 mL/min; Wavelength = 210 nm

Dipeptide	RT	% in the reaction mixture	% Yield	Ratio of dipeptide
Ts-Pro-Pro-OMe	8.85	11	9.4	12
Ts-Pro-Gly-OMe	6.86	48	46	57
Ts-Gly-Pro-OMe	6.39	20	19	24
Ts-Gly-Gly-OMe	5.25	05	5.3	07

XXXXII. Reaction of Proline and Tryptophan with WSCDI in water: (General Procedure-VI) HPLC analysis of the reaction product (Figure.C.I.1d and 2d): Mobile Phase : MeCN:H_2O = 60:40; Flow rate : 0.8 mL/min; Wavelength = 210 nm

Dipeptide	RT	% in the reaction mixture	% Yield	Ratio of dipeptide
Ts-Pro-Pro-OMe	6.24	-	-	-
Ts-Pro-Trp-OMe	9.3	-	-	-
Ts-Trp-Pro-OMe	8.0	15	7.6	100
Ts-Trp-Trp-OMe	10.66	-	-	-

XXXXIII. Reaction of Ts-Pro, Ts-Trp, Pro-OMe and Trp-OMe with WSCDI in water : (General Procedure-VII) HPLC analysis of the reaction product (Figure.C.I.1d and 2d): Mobile Phase : MeCN:H_2O = 60:40; Flow rate : 0.8 mL/min; Wavelength = 210 nm

Dipeptide	RT	% in the reaction mixture	% Yield	Ratio of dipeptide
Ts-Pro-Pro-OMe	6.24	-	-	-
Ts-Pro-Trp-OMe	9.3	39	42	45
Ts-Trp-Pro-OMe	8.0	-	-	-
Ts-Trp-Trp-OMe	10.66	56	51	55

XXXXIV. Reaction of Boc-Pro-Phe, Boc-Phe-Pro, Pro-Phe-OMe and Phe-Pro-OMe with WSCDI in water : (General Procedure-VII) HPLC analysis of the reaction product (Figure.C.I.1e and 2e): Mobile Phase : MeCN:H_2O = 60:40; Flow rate : 0.8 mL/min; Wavelength = 210 nm

Tetrapeptide	RT	% in the reaction mixture	% Yield	Ratio of tetrapeptide
Boc-Pro-Phe-Pro-Phe-OMe	13.68	21	17	23
Boc-Pro-Phe-Phe-Pro-OMe	11.0	69	58	77
Boc-Phe-Pro-Pro-Phe-OMe	10.88	-	-	-
Boc-Phe-Pro-Phe-Pro-OMe	11.4	-	-	-

Preparation of O-Acetyl Gramicidin:

A solution of 200 mg of Gramicidin A in 4 mL of pyridine and 1 mL of acetic anhydride was kept at room temperature for 30 hrs. Ice-cooled water was added to the reaction mixture, compound was precipitated, centrifugated, precipitate washed with cold water (3 x 5 mL). The residue was dissolved in methanol, solvent evaporated to afford 154 mg (75%) O-Acetyl Gramicidin[37].

Oxidation of Acetyl Gramicidin A with RuO_4 :

A suspension of acetyl gramicidin (70 mg, 0.036 mmol) in CCl_4 : CH_3CN : H_2O (1.5 : 1.5 : 3 mL) was admixed with $NaIO_4$ (142 mg, 0.648 mmol, 0.25 eq), $RuCl_3.3H_2O$ (<1 mg), cooled in ice, sealed, left shaken at room temperature for 8 hours, cooled in ice, cautiously opened, filtered, residue washed with CCl_4 (2 mL), CH_3CN (2 mL), the combined filtrate evaporated *in vacuo*, extracted with MeOH (3 x 5 mL) and evaporated. HPLC performed on the residue (0.38 mg) on a reverse phase column [Pep RPC 5/5, Pharmacia, using UV monitor (UV-M, Pharmacia) at 214 nm], using a dual solvent gradient system [solvent A : 0.1% TFA in water; solvent B : 0.1% TFA in acetonitrile; t(min) (A:B%) : 0 (100:0), 5 (80:20), 45 (70:30), 50 (100:0)], led to the isolation of eluents of the two major peaks at retention times, respectively, 25 and 29 minutes (fraction II and fraction I respectively). Amino acid analysis and peptide sequencing was done for both the fractions.

Amino acid analysis of fraction I, at once showed the transformation of one of the four tryptophan residues to aspartic acid. Peptide sequencing on an applied Biosystems 473A sequencer showed that it was Asp^9GA-OAc.

The more polar fraction II, when similarly processed, was found to have the sequence Asp^9Asp^{11}GA-OAc.

F. REFERENCES

[1] von Dohren, H.; Kleinkauf, H. in *The roots of modern biochemistry* (Kleinkauf, H.; von Dohren, H.; Jaenicke, R. eds)1988, pp 355-368, de Gruyter, Berlin, New York.

[2] Lipmann, F. *Science* **1973**, *173*, 875.

[3] Kleinkauf, H.; von Dohren, H. *in Trends in antibiotic research* **1982**, pp 220-232, Japan Antibiotics Research Association, Tokyo.

[4] Lipmann, F. in *The mechanism of enzyme action* (McElroy, W. D.; Glass, B. eds) **1954**, p. 599, Johns Hopkins University Press, Baltimore.

[5] von Dohren, H. *in Biochemistry of peptide antibiotics* (Kleinkauf, H.; von Dohren, H. eds) **1990**, pp 411-507, de Gruyter, Berlin.

[6] Kleinkauf, H.; von Dohren, H. *Eur. J. Biochem.* **1990**, *192*, 1.

[7] Kurahashi, K.; Komura, S.; Akashi, K.; Nishio, C. in *Peptide antibiotics - biosynthesis and functions* (Kleinkauf, H.; von Dohren, H. eds) **1982**, pp 275-288, de Gruyter, Berlin.

[8] Vater, J. in *Biochemistry of peptide antibiotics* (Kleinkauf, H.; von Dohren, H. eds) **1990**, pp 33-55 de Gruyter, Berlin.

[9] Kleinkauf, H. von Dohren, H. (eds) *Peptide antibiotics - biosynthesis and functions* **1982**, de Gruyter, Berlin.

[10] Mukhopadyay, N. K.; Majumder, S.; Ghosh, S. K.; Bose, S. K. *Biochem. J.* **1986**, *240*, 265.

[11] Keller, U.*J. Biol. Chem.* **1987**, *262*, 5852.

[12] Schlumbohm, W.; Keller, U. *Abstracts Workshop Biochemie, Biologie und Molekularbiologie von Streptomyceten* **1989**, p. 36, Technical University, Berlin.

[13] Froyshov, O. *Eur. J. Biochem.* **1975**, *59*, 201.

[14] Lipmann, F. *Acc. Chem. Res.* **1973**, *6*, 361.

[15] Stretton, A. O. W.; Kaplan, S.; Brenner, S. *Cold Spring Harbor Symp. Quant. Biol.* **1966**, *31*, 173.

[16] Roskoski, Jr. R.; Kleinkauf, H.; Gevers, W.; Lipmann, F. *Biochemistry* **1970**, *9*, 4846.

[17] Chou, P. Y.; Fasman, G. D. *Ann. Rev. Biochem.* **1978**, *47*, 251.

[18] Chou, P. Y.; Fasman, G. D. *Adv. Enzymol.* **1978**, *47*, 45.

[19] Argos, P.; Hanei, M.; Garavito, R. M. *FEBS Letters* **1978**, *93*, 18.

[20] Pavletich, N. P.; Pabo, C. O. *Science* **1991**, *252*, 809.

[21] Krizek, B. A.; Amann, B. T.; Kilfoil, V. J.; Merkle, D. L.; Berg, J. M. *J. Am. Chem. Soc.* **1991**, *113*, 4518.

[22] Kochoyan, M.; Keutmann, H.T.; Weiss, M. A. *Biochemistry* **1991**, *30*, 7063.

[23] Bellefroid, E. J.; Lecocq, P. J.; Benhida, A.; Poncelet, D. A.; Belayew, A.; Martial, J. A. *DNA* **1989**, *8*, 377.

[24] Crossley, P. H.; Little, P. F. R. *Proc. Natl. Acad. Sci. USA.* **1991**, *88*, 7923.

[25] Jacobs, G. H. *The EMBO Journal* **1992**, *11*, 4507.

[26] Barton, D. H. R.; Herve, Y.; Potier, P.; Thierny, J. *J. Chem. Soc. Chem. Commun.* **1984**, 1298.

[27] Kaiser, E. T.; Lawrence, D. S. *Science* **1984**, *226*, 505.

[28] Offord, R. E. *Prot. Eng.* **1987**, *1*, 151.

[29] Kaiser, E. T. *Angew. Chem. Int. Ed. Engl.* **1988**, *27*, 913.

[30] West, J. B.; Scholten, J.; Stolowich, N. J.; Hogg, J. L.; Scott, A. I. *J. Am. Chem. Soc.* **1989**, *111*, 4513.

[31] Rana, T. M.; Meares, C. F. *J. Am. Chem. Soc.* **1990**, *112*, 2457.

[32] Veatch, W. R.; Blout, E. R. *Biochemistry* **1974**, *13*, 5257.

[33] Sychev, S. V.; Nevskaya, N. A.; Jordanov, St.; Shepel, E. N.; Miroshnikov, A. T.; Ivanov, V. T. *Bio-org. Chem.* **1980**, *9*,121.

[34] Wallace, B. A.; Ravikumar, K. *Science* **1988**, *241*, 182.

[35] Langs, D. A. *Science* **1988**, *241*, 188.

[36] Langs, D. A.; Smith, G. D.; Courseille, C.; Precigoux, G.; Hospital, M. *Proc. Natl. Acad. Sci. USA.* **1991**, *88*, 5345.

[37] Sarges, R.; Witkop, B. *J. Am. Chem. Soc.* **1965**, *87*, 2011.

[38] Ranganathan, S.; Ranganathan, D.; Bhattacharyya, D. *J. Chem. Soc. Chem. Commun.* **1987**, 1085.

[39] Bystrov, V. F.; Arseniev, A. S. *Tetrahedron* **1988**, *44*, 925.

[40] Curtius, T.; Goebel, F. *J. Prakt. Chem*, **1888**, *37*, 150.

[41] Schott, H. F.; Larkin, J. B.; Rocklandaw, L. B.; M.S.Dunn, M. S. *J. Org. Chem.*, **1947**, *12*, 490.

[42] Jaquenod, P. A.; Waller, J. P. *Helv. Chim. Acta*, **1956**, *39*, 1421.

[43] Erlanger, B. F.; Sachs, H.; Brand, E. *J. Am. Chem. Soc.* **1954**, *76*, 1806.

[44] Guttmann, S.; Boissonnas, R. A. *Helv. Chim. Acta*, **1958**, *41*, 1852.

[45] McChesney, E. W.; Swann, W. K. *J. Am. Chem. Soc.*, **1937**, *59*, 116.

[46] Izumiya, N. *Bull. Chem. Soc. Japan*, **1953**, *26*, 53.

[47] Theodoropoulos, D.; Craig, L. C. *J. Org. Chem*, **1956**, *21*, 1376.

[48] Fischer, E.; Lipschitz, W. *Ber.*, **1915**, *48*, 360.

[49] Oseki, T. *J. Tokyo Chem. Soc.*, **1920**, *41*, 8.

[50] Anderson, G. W.; McGregor, A. C. *J. Am. Chem. Soc.*, **1957**, *79*, 6180.

[51] Kasafirch, E.; Jost, K.; Rudinger, J.; Sorm, F. *Collection Czech. Chem. Commun.* **1965**, *30*, 2600.

[52] Paul, R.; Anderson, G. W. *J. Org. Chem.*, **1962**, *27*, 2094.

APPENDIX.C.I.1

COMPND	ACTINOXANTHIN		
HEADER	ANTIBACTERIAL PROTEIN		
SEQRES	1	108	ALA PRO ALA PHE SER VAL SER PRO ALA SER GLY ALA SER
SEQRES	2	108	ASP GLY GLN SER VAL SER VAL SER VAL ALA ALA ALA GLY
SEQRES	3	108	GLU THR TYR TYR ILE ALA GLN CYS ALA PRO VAL GLY GLY
SEQRES	4	108	GLN ASP ALA CYS ASN PRO ALA THR ALA THR SER PHE THR
SEQRES	5	108	THR ASP ALA SER GLY ALA ALA SER PHE SER PHE THR VAL
SEQRES	6	108	ARG LYS SER TYR ALA GLY GLN THR PRO SER GLY THR PRO
SEQRES	7	108	VAL GLY SER VAL ASP CYS ALA THR ASP ALA CYS ASN LEU
SEQRES	8	108	GLY ALA GLY ASN SER GLY LEU ASN LEU GLY HIS VAL ALA
SEQRES	9	108	LEU THR PHE GLY

SHEET	1	SH1	ALA	3	VAL	6
SHEET	2	SH1	GLN	16	ALA	23
SHEET	3	SH1	SER	60	THR	64
SHEET	1	SH2	GLN	40	CYS	43
SHEET	2	SH2	THR	28	VAL	37
SHEET	3	SH2	CYS	89	GLY	94
SHEET	4	SH2	HIS	102	VAL	103
SHEET	1	SH3	SER	67	THR	72
SHEET	2	SH3	SER	81	ASP	83

COMPND	CALCIUM-BINDING PARVALBUMIN B		
HEADER	CALCIUM BINDING		
SEQRES	1	108	ALA PHE ALA GLY VAL LEU ASN ASP ALA ASP ILE ALA ALA
SEQRES	2	108	ALA LEU GLU ALA CYS LYS ALA ALA ASP SER PHE ASN HIS
SEQRES	3	108	LYS ALA PHE PHE ALA LYS VAL GLY LEU THR SER LYS SER
SEQRES	4	108	ALA ASP ASP VAL LYS LYS ALA PHE ALA ILE ILE ASP GLN
SEQRES	5	108	ASP LYS SER GLY PHE ILE GLU GLU ASP GLU LEU LYS LEU
SEQRES	6	108	PHE LEU GLN ASN PHE LYS ALA ASP ALA ARG ALA LEU THR
SEQRES	7	108	ASP GLY GLU THR LYS THR PHE LEU LYS ALA GLY ASP SER
SEQRES	8	108	ASP GLY ASP GLY LYS ILE GLY VAL ASP GLU PHE THR ALA
SEQRES	9	108	LEU VAL LYS ALA

HELIX	1	A	ASN	7	LEU	15
HELIX	2	B	HIS	26	VAL	33
HELIX	3	C	ALA	40	ASP	51
HELIX	4	D	LEU	67	LYS	71

HELIX	5	E	THR	78	GLY	89
HELIX	6	F	PHE	102	LYS	107

COMPND CRAMBIN
HEADER PLANT SEED PROTEIN
SEQRES 1 46 THR THR CYS CYS PRO SER ILE VAL ALA ARG SER ASN PHE
SEQRES 2 46 ASN VAL CYS ARG LEU PRO GLY THR PRO GLU ALA ILE CYS
SEQRES 3 46 ALA THR TYR THR GLY CYS ILE ILE ILE PRO GLY ALA THR
SEQRES 4 46 CYS PRO GLY ASP TYR ALA ASN

HELIX	1	H1	ILE	7	PRO	19
HELIX	2	H2	GLU	23	THR	30
SHEET	1	S1	THR	1	CYS	4
SHEET	2	S1	CYS	32	ILE	35

COMPND SUBTILISIN CARLSBERG (E.C.3.4.21.14) (COMMERCIAL PRODUCT
COMPND 1 FROM SERRA, HEIDELBERG CALLED SUBTILISIN NAGARSE) COMPLEX
COMPND 2 WITH EGLIN-C
HEADER COMPLEX(SERINE PROTEINASE-INHIBITOR)
SEQRES 1 E 274 ALA GLN THR VAL PRO TYR GLY ILE PRO LEU ILE LYS ALA
SEQRES 2 E 274 ASP LYS VAL GLN ALA GLN GLY PHE LYS GLY ALA ASN VAL
SEQRES 3 E 274 LYS VAL ALA VAL LEU ASP THR GLY ILE GLN ALA SER HIS
SEQRES 4 E 274 PRO ASP LEU ASN VAL VAL GLY GLY ALA SER PHE VAL ALA
SEQRES 5 E 274 GLY GLU ALA TYR ASN THR ASP GLY ASN GLY HIS GLY THR
SEQRES 6 E 274 HIS VAL ALA GLY THR VAL ALA ALA LEU ASP ASN THR THR
SEQRES 7 E 274 GLY VAL LEU GLY VAL ALA PRO SER VAL SER LEU TYR ALA
SEQRES 8 E 274 VAL LYS VAL LEU ASN SER SER GLY SER GLY SER TYR SER
SEQRES 9 E 274 GLY ILE VAL SER GLY ILE GLU TRP ALA THR THR ASN GLY
SEQRES 10 E 274 MET ASP VAL ILE ASN MET SER LEU GLY GLY ALA SER GLY
SEQRES 11 E 274 SER THR ALA MET LYS GLN ALA VAL ASP ASN ALA TYR ALA
SEQRES 12 E 274 ARG GLY VAL VAL VAL VAL ALA ALA ALA GLY ASN SER GLY
SEQRES 13 E 274 ASN SER GLY SER THR ASN THR ILE GLY TYR PRO ALA LYS
SEQRES 14 E 274 TYR ASP SER VAL ILE ALA VAL GLY ALA VAL ASP SER ASN
SEQRES 15 E 274 SER ASN ARG ALA SER PHE SER SER VAL GLY ALA GLU LEU
SEQRES 16 E 274 GLU VAL MET ALA PRO GLY ALA GLY VAL TYR SER THR TYR
SEQRES 17 E 274 PRO THR ASN THR TYR ALA THR LEU ASN GLY THR SER MET
SEQRES 18 E 274 ALA SER PRO HIS VAL ALA GLY ALA ALA ALA LEU ILE LEU
SEQRES 19 E 274 SER LYS HIS PRO ASN LEU SER ALA SER GLN VAL ARG ASN
SEQRES 20 E 274 ARG LEU SER SER THR ALA THR TYR LEU GLY SER SER PHE

```
SEQRES 21 E  274  TYR TYR GLY LYS GLY LEU ILE ASN VAL GLU ALA ALA ALA
SEQRES 22 E  274  GLN
SEQRES  1 I   70  THR GLU PHE GLY SER GLU LEU LYS SER PHE PRO GLU VAL
SEQRES  2 I   70  VAL GLY LYS THR VAL ASP GLN ALA ARG GLU TYR PHE THR
SEQRES  3 I   70  LEU HIS TYR PRO GLN TYR ASN VAL TYR PHE LEU PRO GLU
SEQRES  4 I   70  GLY SER PRO VAL THR LEU ASP LEU ARG ASN TYR ARG VAL
SEQRES  5 I   70  ARG VAL PHE TYR ASN PRO GLY THR ASN VAL VAL ASN HIS
SEQRES  6 I   70  HIS VAL PRO HIS VAL GLY
HELIX    1  AE TYR    6  ILE   11
HELIX    2  BE LYS   12  GLN   19
HELIX    3  CE GLY   63  ALA   74
HELIX    4  DE SER  103  ASN  117
HELIX    5  EE SER  132  GLY  146
HELIX    6  EF THR  220  HIS  238
HELIX    7  GE SER  242  THR  253
HELIX    9  HE ASN  269  ALA  274
HELIX   10  IA PHE   10  VAL   14
HELIX   11  IB THR   17  TYR   29
SHEET    1 S1E  ASN   43  PHE   50
SHEET    2 S1E  SER   89  VAL   95
SHEET    3 S1E  VAL   26  ASP   32
SHEET    4 S1E  ASP  120  MET  124
SHEET    5 S1E  VAL  148  ALA  153
SHEET    6 S1E  ILE  175  VAL  180
SHEET    7 S1E  GLU  197  GLY  202
SHEET    8 S1E  LYS  265  ILE  268
SHEET    1 S1I  LYS    8  PHE   10
SHEET    2 S1I  HIS   65  GLY   70
SHEET    3 S1I  ARG   51  TYR   56
SHEET    4 S1I  ASN   33  LEU   37

COMPND  L7(SLASH)*L12 50 S RIBOSOMAL PROTEIN (C-TERMINAL DOMAIN)
HEADER  RIBOSOMAL PROTEIN
SEQRES  1   74  ALA ALA GLU GLU LYS THR GLU PHE ASP VAL ILE LEU LYS
SEQRES  2   74  ALA ALA GLY ALA ASN LYS VAL ALA VAL ILE LYS ALA VAL
SEQRES  3   74  ARG GLY ALA THR GLY LEU GLY LEU LYS GLU ALA LYS ASP
SEQRES  4   74  LEU VAL GLU SER ALA PRO ALA ALA LEU LYS GLU GLY VAL
SEQRES  5   74  SER LYS ASP ASP ALA GLU ALA LEU LYS LYS ALA LEU GLU
```

```
SEQRES  6   74  GLU ALA GLY ALA GLU VAL GLU VAL LYS

COMPND   HEMOGLOBIN (ERYTHROCRUORIN, AQUO MET)
HEADER   OXYGEN TRANSPORT
SEQRES  1  136  LEU SER ALA ASP GLN ILE SER THR VAL GLN ALA SER PHE
SEQRES  2  136  ASP LYS VAL LYS GLY ASP PRO VAL GLY ILE LEU TYR ALA
SEQRES  3  136  VAL PHE LYS ALA ASP PRO SER ILE MET ALA LYS PHE THR
SEQRES  4  136  GLN PHE ALA GLY LYS ASP LEU GLU SER ILE LYS GLY THR
SEQRES  5  136  ALA PRO PHE GLU THR HIS ALA ASN ARG ILE VAL GLY PHE
SEQRES  6  136  PHE SER LYS ILE ILE GLY GLU LEU PRO ASN ILE GLU ALA
SEQRES  7  136  ASP VAL ASN THR PHE VAL ALA SER HIS LYS PRO ARG GLY
SEQRES  8  136  VAL THR HIS ASP GLN LEU ASN ASN PHE ARG ALA GLY PHE
SEQRES  9  136  VAL SER TYR MET LYS ALA HIS THR ASP PHE ALA GLY ALA
SEQRES 10  136  GLU ALA ALA TRP GLY ALA THR LEU ASP THR PHE PHE GLY
SEQRES 11  136  MET ILE PHE SER LYS MET
HELIX   1    A SER     2  LYS     17
HELIX   2    B ASP    19  ASP     31
HELIX   3    C ASP    31  PHE     38
HELIX   4    D ASP    45  LYS     50
HELIX   5    E THR    52  GLU     72
HELIX   6    F ILE    76  LYS     88
HELIX   7   FG HIS    87  GLY     91
HELIX   8    G THR    93  THR    112
HELIX   9    H ALA   117  PHE    133

COMPND   IMMUNOGLOBULIN FAB
HEADER   IMMUNOGLOBULIN
SEQRES  1 L  216  GLU SER VAL LEU THR GLN PRO PRO SER ALA SER GLY THR
SEQRES  2 L  216  PRO GLY GLN ARG VAL THR ILE SER CYS THR GLY THR SER
SEQRES  3 L  216  SER ASN ILE GLY SER ILE THR VAL ASN TRP TYR GLN GLN
SEQRES  4 L  216  LEU PRO GLY MET ALA PRO LYS LEU LEU ILE TYR ARG ASP
SEQRES  5 L  216  ALA MET ARG PRO SER GLY VAL PRO THR ARG PHE SER GLY
SEQRES  6 L  216  SER LYS SER GLY THR SER ALA SER LEU ALA ILE SER GLY
SEQRES  7 L  216  LEU GLU ALA GLU ASP GLU SER ASP TYR TYR CYS ALA SER
SEQRES  8 L  216  TRP ASN SER SER ASP ASN SER TYR VAL PHE GLY THR GLY
SEQRES  9 L  216  THR LYS VAL THR VAL LEU GLY GLN PRO LYS ALA ASN PRO
SEQRES 10 L  216  THR VAL THR LEU PHE PRO PRO SER SER GLU GLU LEU GLN
SEQRES 11 L  216  ALA ASN LYS ALA THR LEU VAL CYS LEU ILE SER ASP PHE
```

```
SEQRES 12 L  216  TYR PRO GLY ALA VAL THR VAL ALA TRP LYS ALA ASP GLY
SEQRES 13 L  216  SER PRO VAL LYS ALA GLY VAL GLU THR THR LYS PRO SER
SEQRES 14 L  216  LYS GLN SER ASN ASN LYS TYR ALA ALA SER SER TYR LEU
SEQRES 15 L  216  SER LEU THR PRO GLU GLN TRP LYS SER HIS ARG SER TYR
SEQRES 16 L  216  SER CYS GLN VAL THR HIS GLU GLY SER THR VAL GLU LYS
SEQRES 17 L  216  THR VAL ALA PRO THR GLU CYS SER
SEQRES  1 H  229  GLU VAL GLN LEU VAL GLN SER GLY GLY GLY VAL VAL GLN
SEQRES  2 H  229  PRO GLY ARG SER LEU ARG LEU SER CYS SER SER SER GLY
SEQRES  3 H  229  PHE ILE PHE SER SER TYR ALA MET TYR TRP VAL ARG GLN
SEQRES  4 H  229  ALA PRO GLY LYS GLY LEU GLU TRP VAL ALA ILE ILE TRP
SEQRES  5 H  229  ASP ASP GLY SER ASP GLN HIS TYR ALA ASP SER VAL LYS
SEQRES  6 H  229  GLY ARG PHE THR ILE SER ARG ASN ASP SER LYS ASN THR
SEQRES  7 H  229  LEU PHE LEU GLN MET ASP SER LEU ARG PRO GLU ASP THR
SEQRES  8 H  229  GLY VAL TYR PHE CYS ALA ARG ASP GLY GLY HIS GLY PHE
SEQRES  9 H  229  CYS SER SER ALA SER CYS PHE GLY PRO ASP TYR TRP GLY
SEQRES 10 H  229  GLN GLY THR PRO VAL THR VAL SER SER ALA SER THR LYS
SEQRES 11 H  229  GLY PRO SER VAL PHE PRO LEU ALA PRO SER SER LYS SER
SEQRES 12 H  229  THR SER GLY GLY THR ALA ALA LEU GLY CYS LEU VAL LYS
SEQRES 13 H  229  ASP TYR PHE PRO GLN PRO VAL THR VAL SER TRP ASN SER
SEQRES 14 H  229  GLY ALA LEU THR SER GLY VAL HIS THR PHE PRO ALA VAL
SEQRES 15 H  229  LEU GLN SER SER GLY LEU TYR SER LEU SER SER VAL VAL
SEQRES 16 H  229  THR VAL PRO SER SER SER LEU GLY THR GLN THR TYR ILE
SEQRES 17 H  229  CYS ASN VAL ASN HIS LYS PRO SER ASN THR LYS VAL ASP
SEQRES 18 H  229  LYS ARG VAL GLU PRO LYS SER CYS

COMPND   FERREDOXIN
HEADER   ELECTRON TRANSPORT
SEQRES  1   54  ALA TYR VAL ILE ASN ASP SER CYS ILE ALA CYS GLY ALA
SEQRES  2   54  CYS LYS PRO GLU CYS PRO VAL ASN ILE ILE GLN GLY SER
SEQRES  3   54  ILE TYR ALA ILE ASP ALA ASP SER CYS ILE ASP CYS GLY
SEQRES  4   54  SER CYS ALA SER VAL CYS PRO VAL GLY ALA PRO ASN PRO
SEQRES  5   54  GLU ASP

COMPND   FLAVODOXIN
HEADER   ELECTRON TRANSFER (FLAVOPROTEIN)
SEQRES  1  148  MET PRO LYS ALA LEU ILE VAL TYR GLY SER THR THR GLY
SEQRES  2  148  ASN THR GLU TYR THR ALA GLU THR ILE ALA ARG GLN LEU
SEQRES  3  148  ALA ASN ALA GLY TYR GLU VAL ASP SER ARG ASP ALA ALA
```

```
SEQRES  4  148  SER VAL GLU ALA GLY GLY LEU PHE GLU GLY PHE ASP LEU
SEQRES  5  148  VAL LEU LEU GLY CYS SER THR TRP GLY ASP ASP SER ILE
SEQRES  6  148  GLU LEU GLN ASP ASP PHE ILE PRO LEU PHE ASP SER LEU
SEQRES  7  148  GLU GLU THR GLY ALA GLN GLY ARG LYS VAL ALA CYS PHE
SEQRES  8  148  GLY CYS GLY ASP SER SER TYR GLU TYR PHE CYS GLY ALA
SEQRES  9  148  VAL ASP ALA ILE GLU GLU LYS LEU LYS ASN LEU GLY ALA
SEQRES 10  148  GLU ILE VAL GLN ASP GLY LEU ARG ILE ASP GLY ASP PRO
SEQRES 11  148  ARG ALA ALA ARG ASP ASP ILE VAL GLY TRP ALA HIS ASP
SEQRES 12  148  VAL ARG GLY ALA ILE

COMPND  GAMMA-/II$ CRYSTALLIN
HEADER  CRYSTALLIN
SEQRES  1  174  GLY LYS ILE THR PHE TYR GLU ASP ARG GLY PHE GLN GLY
SEQRES  2  174  HIS CYS TYR GLU CYS SER SER ASP CYS PRO ASN LEU GLN
SEQRES  3  174  PRO TYR PHE SER ARG CYS ASN SER ILE ARG VAL ASP SER
SEQRES  4  174  GLY CYS TRP MET LEU TYR GLU ARG PRO ASN TYR GLN GLY
SEQRES  5  174  HIS GLN TYR PHE LEU ARG ARG GLY ASP TYR PRO ASP TYR
SEQRES  6  174  GLN GLN TRP MET GLY PHE ASN ASP SER ILE ARG SER CYS
SEQRES  7  174  ARG LEU ILE PRO GLN HIS THR GLY THR PHE ARG MET ARG
SEQRES  8  174  ILE TYR GLU ARG ASP ASP PHE ARG GLY GLN MET SER GLU
SEQRES  9  174  ILE THR ASP ASP CYS PRO SER LEU GLN ASP ARG PHE HIS
SEQRES 10  174  LEU SER GLU VAL HIS SER LEU ASN VAL LEU GLU GLY SER
SEQRES 11  174  TRP VAL LEU TYR GLU MET PRO SER TYR ARG GLY ARG GLN
SEQRES 12  174  TYR LEU LEU ARG PRO GLY GLU TYR ARG ARG TYR LEU ASP
SEQRES 13  174  TRP GLY ALA MET ASN ALA LYS VAL GLY SER LEU ARG ARG
SEQRES 14  174  VAL MET ASP PHE TYR
HELIX   1  H1 ASP    64  MET    69
HELIX   2  H2 SER   111  HIS   117
HELIX   3  H3 ARG   153  GLY   158
SHEET   1   A  GLN   12  CYS    18
SHEET   2   A  LYS    2  ASP     8
SHEET   3   A  SER   34  SER    39
SHEET   4   A  GLY   60  TYR    62
SHEET   1   B  ASP   21  PRO    23
SHEET   2   B  SER   77  ILE    81
SHEET   3   B  CYS   41  ARG    47
SHEET   4   B  GLN   51  LEU    57
SHEET   1   C  MET  102  ILE   105
```

```
SHEET    2    C   ARG    89  TYR    93
SHEET    3    C   SER   123  GLU   128
SHEET    4    C   GLY   149  TYR   151
SHEET    1   D1   ASP   108  PRO   110
SHEET    2   D1   ALA   162  ARG   169
SHEET    3   D1   TRP   131  MET   136
SHEET    4   D1   ARG   140  LEU   146
SHEET    1   D2   SER   119  VAL   121

COMPND   OXIDIZED HIGH POTENTIAL IRON PROTEIN (HIPIP).
HEADER   ELECTRON TRANSFER (IRON-SULFUR PROTEIN)
SEQRES   1   85   SER ALA PRO ALA ASN ALA VAL ALA ALA ASP ASN ALA THR
SEQRES   2   85   ALA ILE ALA LEU LYS TYR ASN GLN ASP ALA THR LYS SER
SEQRES   3   85   GLU ARG VAL ALA ALA ALA ARG PRO GLY LEU PRO PRO GLU
SEQRES   4   85   GLU GLN HIS CYS ALA ASP CYS GLN PHE MET GLN ALA ASP
SEQRES   5   85   ALA ALA GLY ALA THR ASP GLU TRP LYS GLY CYS GLN LEU
SEQRES   6   85   PHE PRO GLY LYS LEU ILE ASN VAL ASN GLY TRP CYS ALA
SEQRES   7   85   SER TRP THR LEU LYS ALA GLY
HELIX    1   H1   ALA    12  ALA    16
HELIX    2   H2   ARG    28  ALA    31
SHEET    1   S1   PHE    48  GLN    50
SHEET    2   S1   GLU    59  GLN    64
SHEET    3   S1   LYS    69  VAL    73

COMPND   HEMERYTHRIN (MET)
HEADER   OXYGEN TRANSPORT
SEQRES   1  113   GLY PHE PRO ILE PRO ASP PRO TYR CYS TRP ASP ILE SER
SEQRES   2  113   PHE ARG THR PHE TYR THR ILE VAL ASP ASP GLU HIS LYS
SEQRES   3  113   THR LEU PHE ASN GLY ILE LEU LEU LEU SER GLN ALA ASP
SEQRES   4  113   ASN ALA ASP HIS LEU ASN GLU LEU ARG ARG CYS THR GLY
SEQRES   5  113   LYS HIS PHE LEU ASN GLU GLN GLN LEU MET GLN ALA SER
SEQRES   6  113   GLN TYR ALA GLY TYR ALA GLU HIS LYS LYS ALA HIS ASP
SEQRES   7  113   ASP PHE ILE HIS LYS LEU ASP THR TRP ASP GLY ASP VAL
SEQRES   8  113   THR TYR ALA LYS ASN TRP LEU VAL ASN HIS ILE LYS THR
SEQRES   9  113   ILE ASP PHE LYS TYR ARG GLY LYS ILE
HELIX    1   H1 THR    19  GLN    37
HELIX    2   H2 ALA    41  ALA    64
HELIX    3   H3 TYR    70  ASP    85
```

```
HELIX    4  H4 VAL    91  TYR   109

COMPND   INSULIN
HEADER   HORMONE
SEQRES   1 A  21   GLY ILE VAL GLU GLN CYS CYS THR SER ILE CYS SER LEU
SEQRES   2 A  21   TYR GLN LEU GLU ASN TYR CYS ASN
SEQRES   1 B  30   PHE VAL ASN GLN HIS LEU CYS GLY SER HIS LEU VAL GLU
SEQRES   2 B  30   ALA LEU TYR LEU VAL CYS GLY GLU ARG GLY PHE PHE TYR
SEQRES   3 B  30   THR PRO LYS ALA
HELIX    1 A11 GLY      1  ILE    10
HELIX    2 A12 SER     12  GLU    17
HELIX    3 B11 SER      9  GLY    20
SHEET    1  B  PHE    24  TYR    26

COMPND   MYOGLOBIN (DEOXY, $P*H 8.4)
HEADER   OXYGEN STORAGE
SEQRES   1   153   VAL LEU SER GLU GLY GLU TRP GLN LEU VAL LEU HIS VAL
SEQRES   2   153   TRP ALA LYS VAL GLU ALA ASP VAL ALA GLY HIS GLY GLN
SEQRES   3   153   ASP ILE LEU ILE ARG LEU PHE LYS SER HIS PRO GLU THR
SEQRES   4   153   LEU GLU LYS PHE ASP ARG PHE LYS HIS LEU LYS THR GLU
SEQRES   5   153   ALA GLU MET LYS ALA SER GLU ASP LEU LYS LYS HIS GLY
SEQRES   6   153   VAL THR VAL LEU THR ALA LEU GLY ALA ILE LEU LYS LYS
SEQRES   7   153   LYS GLY HIS HIS GLU ALA GLU LEU LYS PRO LEU ALA GLN
SEQRES   8   153   SER HIS ALA THR LYS HIS LYS ILE PRO ILE LYS TYR LEU
SEQRES   9   153   GLU PHE ILE SER GLU ALA ILE ILE HIS VAL LEU HIS SER
SEQRES  10   153   ARG HIS PRO GLY ASP PHE GLY ALA ASP ALA GLN GLY ALA
SEQRES  11   153   MET ASN LYS ALA LEU GLU LEU PHE ARG LYS ASP ILE ALA
SEQRES  12   153   ALA LYS TYR LYS GLU LEU GLY TYR GLN GLY
HELIX    1    A SER      3  GLU      18
HELIX    2    B ASP     20  SER      35
HELIX    3    C HIS     36  LYS      42
HELIX    4    D THR     51  ALA      57
HELIX    5    E SER     58  LYS      77
HELIX    6    F LEU     86  THR      95
HELIX    7    G PRO    100  ARG     118
HELIX    8    H GLY    124  LEU     149
```

```
COMPND  MELITTIN
HEADER  TOXIN (HEMOLYTIC POLYPEPTIDE)
SEQRES  1 A  26  GLY ILE GLY ALA VAL LEU LYS VAL LEU THR THR GLY LEU
SEQRES  2 A  26  PRO ALA LEU ILE SER TRP ILE LYS ARG LYS ARG GLN GLN
HELIX   1   GLY    1  THR   10
HELIX   2   LEU   13  GLN   26

COMPND  PLASTOCYANIN (CU ++, $P*H 6.0)
HEADER  ELECTRON TRANSPORT (CU BINDING PROTEIN)
SEQRES  1   99  ILE ASP VAL LEU LEU GLY ALA ASP ASP GLY SER LEU ALA
SEQRES  2   99  PHE VAL PRO SER GLU PHE SER ILE SER PRO GLY GLU LYS
SEQRES  3   99  ILE VAL PHE LYS ASN ASN ALA GLY PHE PRO HIS ASN ILE
SEQRES  4   99  VAL PHE ASP GLU ASP SER ILE PRO SER GLY VAL ASP ALA
SEQRES  5   99  SER LYS ILE SER MET SER GLU GLU ASP LEU LEU ASN ALA
SEQRES  6   99  LYS GLY GLU THR PHE GLU VAL ALA LEU SER ASN LYS GLY
SEQRES  7   99  GLU TYR SER PHE TYR CYS SER PRO HIS GLN GLY ALA GLY
SEQRES  8   99  MET VAL GLY LYS VAL THR VAL ASN
HELIX   1  A  ALA   52  SER   56
SHEET   1  S  MET   57  LEU   63
SHEET   2  S  HIS   37  SER   45
SHEET   3  S  GLY   78  HIS   87
SHEET   4  S  MET   92  ASN   99
SHEET   5  S  SER   11  ILE   21
SHEET   6  S  ILE    1  ALA    7
SHEET   7  S  LYS   26  ALA   33
SHEET   8  S  GLY   67  LEU   74

COMPND  AVIAN PANCREATIC POLYPEPTIDE
HEADER  PANCREATIC HORMONE
SEQRES  1   36  GLY PRO SER GLN PRO THR TYR PRO GLY ASP ASP ALA PRO
SEQRES  2   36  VAL GLU ASP LEU ILE ARG PHE TYR ASP ASN LEU GLN GLN
SEQRES  3   36  TYR LEU ASN VAL VAL THR ARG HIS ARG TYR
HELIX   1  A PRO    2  PRO    8
HELIX   2  B VAL   14  THR   32

COMPND  BENCE-*JONES IMMUNOGLOBULIN /REI$ VARIABLE PORTION
HEADER  IMMUNOGLOBULIN(PART)SEQUESTERS ANTIGENS
```

```
SEQRES  1  107  ASP ILE GLN MET THR GLN SER PRO SER SER LEU SER ALA
SEQRES  2  107  SER VAL GLY ASP ARG VAL THR ILE THR CYS GLN ALA SER
SEQRES  3  107  GLN ASP ILE ILE LYS TYR LEU ASN TRP TYR GLN GLN THR
SEQRES  4  107  PRO GLY LYS ALA PRO LYS LEU LEU ILE TYR GLU ALA SER
SEQRES  5  107  ASN LEU GLN ALA GLY VAL PRO SER ARG PHE SER GLY SER
SEQRES  6  107  GLY SER GLY THR ASP TYR THR PHE THR ILE SER SER LEU
SEQRES  7  107  GLN PRO GLU ASP ILE ALA THR TYR TYR CYS GLN GLN TYR
SEQRES  8  107  GLN SER LEU PRO TYR THR PHE GLY GLN GLY THR LYS LEU
SEQRES  9  107  GLN ILE THR
SHEET   1  AA  THR    5  SER   7
SHEET   2  AA  VAL   19  GLN  24
SHEET   3  AA  TYR   71  ILE  75
SHEET   4  AA  SER   63  SER  65
SHEET   1  AB  SER    9  ALA  13
SHEET   2  AB  GLY   99  THR  107
SHEET   3  AB  THR   85  GLN   89
SHEET   4  AB  ASN   34  GLN   38
SHEET   5  AB  ALA   43  GLU   50

COMPND  RIBONUCLEASE T=1=(E.C.3.1.27.3) ISOZYME-2(PRIME)-GUANYLIC
COMPND  2 ACID COMPLEX
HEADER  HYDROLASE(ENDORIBONUCLEASE)
SEQRES  1  104  ALA CYS ASP TYR THR CYS GLY SER ASN CYS TYR SER SER
SEQRES  2  104  SER ASP VAL SER THR ALA GLN ALA ALA GLY TYR LYS LEU
SEQRES  3  104  HIS GLU ASP GLY GLU THR VAL GLY SER ASN SER TYR PRO
SEQRES  4  104  HIS LYS TYR ASN ASN TYR GLU GLY PHE ASP PHE SER VAL
SEQRES  5  104  SER SER PRO TYR TYR GLU TRP PRO ILE LEU SER SER GLY
SEQRES  6  104  ASP VAL TYR SER GLY GLY SER PRO GLY ALA ASP ARG VAL
SEQRES  7  104  VAL PHE ASN GLU ASN ASN GLN LEU ALA GLY VAL ILE THR
SEQRES  8  104  HIS THR GLY ALA SER GLY ASN ASN PHE VAL GLU CYS THR
HELIX   1   A  SER    13  ASP     29
SHEET   1  S1  TYR     4  CYS      6
SHEET   2  S1  ASN     9  SER     12
SHEET   1  S2  PRO    39  TYR     42
SHEET   2  S2  PRO    55  LEU     62
SHEET   3  S2  ASP    76  ASN     81
SHEET   4  S2  GLN    85  THR     91
SHEET   5  S2  PHE   100  CYS    103
```

```
COMPND  SCORPION NEUROTOXIN (VARIANT 3)
HEADER  TOXIN
SEQRES  1   65  LYS GLU GLY TYR LEU VAL LYS LYS SER ASP GLY CYS LYS
SEQRES  2   65  TYR GLY CYS LEU LYS LEU GLY GLU ASN GLU GLY CYS ASP
SEQRES  3   65  THR GLU CYS LYS ALA LYS ASN GLN GLY GLY SER TYR GLY
SEQRES  4   65  TYR CYS TYR ALA PHE ALA CYS TRP CYS GLU GLY LEU PRO
SEQRES  5   65  GLU SER THR PRO THR TYR PRO LEU PRO ASN LYS SER CYS
HELIX   1   H1  GLU   23  LYS   32

COMPND  TONIN (E.C. NUMBER NOT ASSIGNED)
HEADER  HYDROLASE(SERINE PROTEINASE)
SEQRES  1  235  ILE VAL GLY GLY TYR LYS CYS GLU LYS ASN SER GLN PRO
SEQRES  2  235  TRP GLN VAL ALA VAL ILE ASN GLU TYR LEU CYS GLY GLY
SEQRES  3  235  VAL LEU ILE ASP PRO SER TRP VAL ILE THR ALA ALA HIS
SEQRES  4  235  CYS TYR SER ASN ASN TYR GLN VAL LEU LEU GLY ARG ASN
SEQRES  5  235  ASN LEU PHE LYS ASP GLU PRO PHE ALA GLN ARG ARG LEU
SEQRES  6  235  VAL ARG GLN SER PHE ARG HIS PRO ASP TYR ILE PRO LEU
SEQRES  7  235  ILE VAL THR ASN ASP THR GLU GLN PRO VAL HIS ASP HIS
SEQRES  8  235  SER ASN ASP LEU MET LEU LEU HIS LEU SER GLU PRO ALA
SEQRES  9  235  ASP ILE THR GLY GLY VAL LYS VAL ILE ASP LEU PRO THR
SEQRES 10  235  LYS GLU PRO LYS VAL GLY SER THR CYS LEU ALA SER GLY
SEQRES 11  235  TRP GLY SER THR ASN PRO SER GLU MET VAL VAL SER HIS
SEQRES 12  235  ASP LEU GLN CYS VAL ASN ILE HIS LEU LEU SER ASN GLU
SEQRES 13  235  LYS CYS ILE GLU THR TYR LYS ASP ASN VAL THR ASP VAL
SEQRES 14  235  MET LEU CYS ALA GLY GLU MET GLU GLY GLY LYS ASP THR
SEQRES 15  235  CYS ALA GLY ASP SER GLY GLY PRO LEU ILE CYS ASP GLY
SEQRES 16  235  VAL LEU GLN GLY ILE THR SER GLY GLY ALA THR PRO CYS
SEQRES 17  235  ALA LYS PRO LYS THR PRO ALA ILE TYR ALA LYS LEU ILE
SEQRES 18  235  LYS PHE THR SER TRP ILE LYS LYS VAL MET LYS GLU ASN
SEQRES 19  235  PRO

COMPND  BETA-TRYPSIN (E.C.3.4.21.4) COMPLEX WITH
COMPND  2 P-AMIDINO-PHENYL-PYRUVATE (APPA)
HEADER  HYDROLASE (SERINE PROTEINASE)
SEQRES  1  223  ILE VAL GLY GLY TYR THR CYS GLY ALA ASN THR VAL PRO
SEQRES  2  223  TYR GLN VAL SER LEU ASN SER GLY TYR HIS PHE CYS GLY
SEQRES  3  223  GLY SER LEU ILE ASN SER GLN TRP VAL VAL SER ALA ALA
```

```
SEQRES  4  223  HIS CYS TYR LYS SER GLY ILE GLN VAL ARG LEU GLY GLU
SEQRES  5  223  ASP ASN ILE ASN VAL VAL GLU GLY ASN GLU GLN PHE ILE
SEQRES  6  223  SER ALA SER LYS SER ILE VAL HIS PRO SER TYR ASN SER
SEQRES  7  223  ASN THR LEU ASN ASN ASP ILE MET LEU ILE LYS LEU LYS
SEQRES  8  223  SER ALA ALA SER LEU ASN SER ARG VAL ALA SER ILE SER
SEQRES  9  223  LEU PRO THR SER CYS ALA SER ALA GLY THR GLN CYS LEU
SEQRES 10  223  ILE SER GLY TRP GLY ASN THR LYS SER SER GLY THR SER
SEQRES 11  223  TYR PRO ASP VAL LEU LYS CYS LEU LYS ALA PRO ILE LEU
SEQRES 12  223  SER ASP SER SER CYS LYS SER ALA TYR PRO GLY GLN ILE
SEQRES 13  223  THR SER ASN MET PHE CYS ALA GLY TYR LEU GLU GLY GLY
SEQRES 14  223  LYS ASP SER CYS GLN GLY ASP SER GLY GLY PRO VAL VAL
SEQRES 15  223  CYS SER GLY LYS LEU GLN GLY ILE VAL SER TRP GLY SER
SEQRES 16  223  GLY CYS ALA GLN LYS ASN LYS PRO GLY VAL TYR THR LYS
SEQRES 17  223  VAL CYS ASN TYR VAL SER TRP ILE LYS GLN THR ILE ALA
SEQRES 18  223  SER ASN

COMPND  UBIQUITIN
HEADER  CHROMOSOMAL PROTEIN
SEQRES  1   76  MET GLN ILE PHE VAL LYS THR LEU THR GLY LYS THR ILE
SEQRES  2   76  THR LEU GLU VAL GLU PRO SER ASP THR ILE GLU ASN VAL
SEQRES  3   76  LYS ALA LYS ILE GLN ASP LYS GLU GLY ILE PRO PRO ASP
SEQRES  4   76  GLN GLN ARG LEU ILE PHE ALA GLY LYS GLN LEU GLU ASP
SEQRES  5   76  GLY ARG THR LEU SER ASP TYR ASN ILE GLN LYS GLU SER
SEQRES  6   76  THR LEU HIS LEU VAL LEU ARG LEU ARG GLY GLY
HELIX   1  H1   ILE    23  GLU    34
HELIX   2  H2   LEU    56  TYR    59
SHEET   1  BET  GLY    10  VAL    17
SHEET   2  BET  MET     1  THR     7
SHEET   3  BET  GLU    64  ARG    72
SHEET   4  BET  GLN    40  PHE    45
SHEET   5  BET  LYS    48  LEU    50

COMPND  ACTINIDIN (SULFHYDRYL PROTEINASE) (E.C. NUMBER NOT ASSIGNED)
HEADER  HYDROLASE (PROTEINASE)
SEQRES  1  220  LEU PRO SER TYR VAL ASP TRP ARG SER ALA GLY ALA VAL
SEQRES  2  220  VAL ASP ILE LYS SER GLN GLY GLU CYS GLY GLY CYS TRP
SEQRES  3  220  ALA PHE SER ALA ILE ALA THR VAL GLU GLY ILE ASN LYS
SEQRES  4  220  ILE THR SER GLY SER LEU ILE SER LEU SER GLU GLN GLU
```

```
SEQRES  5  220  LEU ILE ASP CYS GLY ARG THR GLN ASN THR ARG GLY CYS
SEQRES  6  220  ASP GLY GLY TYR ILE THR ASP GLY PHE GLN PHE ILE ILE
SEQRES  7  220  ASN ASP GLY GLY ILE ASN THR GLU GLU ASN TYR PRO TYR
SEQRES  8  220  THR ALA GLN ASP GLY ASP CYS ASP VAL ALA LEU GLN ASP
SEQRES  9  220  GLN LYS TYR VAL THR ILE ASP THR TYR GLU ASN VAL PRO
SEQRES 10  220  TYR ASN ASN GLU TRP ALA LEU GLN THR ALA VAL THR TYR
SEQRES 11  220  GLN PRO VAL SER VAL ALA LEU ASP ALA ALA GLY ASP ALA
SEQRES 12  220  PHE LYS GLN TYR ALA SER GLY ILE PHE THR GLY PRO CYS
SEQRES 13  220  GLY THR ALA VAL ASP HIS ALA ILE VAL ILE VAL GLY TYR
SEQRES 14  220  GLY THR GLU GLY GLY VAL ASP TYR TRP ILE VAL LYS ASN
SEQRES 15  220  SER TRP ASP THR THR TRP GLY GLU GLU GLY TYR MET ARG
SEQRES 16  220  ILE LEU ARG ASN VAL GLY GLY ALA GLY THR CYS GLY ILE
SEQRES 17  220  ALA THR MET PRO SER TYR PRO VAL LYS TYR ASN ASN
HELIX    1  A1 GLY    24  GLY    43
HELIX    2  A2 GLU    50  GLY    57
HELIX    3  A3 TYR    69  GLY    81
HELIX    4  A4 ASP    99  ASP   104
HELIX    5  A5 ASN   120  TYR   130
HELIX    6  A6 GLY   141  TYR   147
SHEET    1  B1  VAL    5  TRP     7
SHEET    2  B1  HIS  162  GLU   172
SHEET    3  B1  VAL  175  LYS   181
SHEET    4  B1  TYR  193  ARG   198
SHEET    5  B1  PHE  152  PHE   152
SHEET    1  B2  VAL  133  LEU   137
SHEET    2  B2  HIS  162  GLU   172
SHEET    3  B2  VAL  175  LYS   181
SHEET    4  B2  TYR  193  ARG   198
SHEET    5  B2  PHE  152  PHE   152

COMPND  ALPHA-LYTIC PROTEASE (E.C. NUMBER NOT ASSIGNED)
HEADER  HYDROLASE (SERINE PROTEINASE)
SEQRES  1  198  ALA ASN ILE VAL GLY GLY ILE GLU TYR SER ILE ASN ASN
SEQRES  2  198  ALA SER LEU CYS SER VAL GLY PHE SER VAL THR ARG GLY
SEQRES  3  198  ALA THR LYS GLY PHE VAL THR ALA GLY HIS CYS GLY THR
SEQRES  4  198  VAL ASN ALA THR ALA ARG ILE GLY GLY ALA VAL VAL GLY
SEQRES  5  198  THR PHE ALA ALA ARG VAL PHE PRO GLY ASN ASP ARG ALA
SEQRES  6  198  TRP VAL SER LEU THR SER ALA GLN THR LEU LEU PRO ARG
```

```
SEQRES  7  198  VAL ALA ASN GLY SER SER PHE VAL THR VAL ARG GLY SER
SEQRES  8  198  THR GLU ALA ALA VAL GLY ALA ALA VAL CYS ARG SER GLY
SEQRES  9  198  ARG THR THR GLY TYR GLN CYS GLY THR ILE THR ALA LYS
SEQRES 10  198  ASN VAL THR ALA ASN TYR ALA GLU GLY ALA VAL ARG GLY
SEQRES 11  198  LEU THR GLN GLY ASN ALA CYS MET GLY ARG GLY ASP SER
SEQRES 12  198  GLY GLY SER TRP ILE THR SER ALA GLY GLN ALA GLN GLY
SEQRES 13  198  VAL MET SER GLY GLY ASN VAL GLN SER ASN GLY ASN ASN
SEQRES 14  198  CYS GLY ILE PRO ALA SER GLN ARG SER SER LEU PHE GLU
SEQRES 15  198  ARG LEU GLN PRO ILE LEU SER GLN TYR GLY LEU SER LEU
SEQRES 16  198  VAL THR GLY

COMPND   ACID PROTEINASE (E.C.3.4.23.7),PENICILLOPEPSIN
HEADER   HYDROLASE (ACID PROTEINASE)
SEQRES  1  323  ALA ALA SER GLY VAL ALA THR ASN THR PRO THR ALA ASN
SEQRES  2  323  ASP GLU GLU TYR ILE THR PRO VAL THR ILE GLY GLY THR
SEQRES  3  323  THR LEU ASN LEU ASN PHE ASP THR GLY SER ALA ASP LEU
SEQRES  4  323  TRP VAL PHE SER THR GLU LEU PRO ALA SER GLN GLN SER
SEQRES  5  323  GLY HIS SER VAL TYR ASN PRO SER ALA THR GLY LYS GLU
SEQRES  6  323  LEU SER GLY TYR THR TRP SER ILE SER TYR GLY ASP GLY
SEQRES  7  323  SER SER ALA SER GLY ASN VAL PHE THR ASP SER VAL THR
SEQRES  8  323  VAL GLY GLY VAL THR ALA HIS GLY GLN ALA VAL GLN ALA
SEQRES  9  323  ALA GLN GLN ILE SER ALA GLN PHE GLN GLN ASP THR ASN
SEQRES 10  323  ASN ASP GLY LEU LEU GLY LEU ALA PHE SER SER ILE ASN
SEQRES 11  323  THR VAL GLN PRO GLN SER GLN THR THR PHE PHE ASP THR
SEQRES 12  323  VAL LYS SER SER LEU ALA GLN PRO LEU PHE ALA VAL ALA
SEQRES 13  323  LEU LYS HIS GLN GLN PRO GLY VAL TYR ASP PHE GLY PHE
SEQRES 14  323  ILE ASP SER SER LYS TYR THR GLY SER LEU THR TYR THR
SEQRES 15  323  GLY VAL ASP ASN SER GLN GLY PHE TRP SER PHE ASN VAL
SEQRES 16  323  ASP SER TYR THR ALA GLY SER GLN SER GLY ASP GLY PHE
SEQRES 17  323  SER GLY ILE ALA ASP THR GLY THR THR LEU LEU LEU LEU
SEQRES 18  323  ASP ASP SER VAL VAL SER GLN TYR TYR SER GLN VAL SER
SEQRES 19  323  GLY ALA GLN GLN ASP SER ASN ALA GLY GLY TYR VAL PHE
SEQRES 20  323  ASP CYS SER THR ASN LEU PRO ASP PHE SER VAL SER ILE
SEQRES 21  323  SER GLY TYR THR ALA THR VAL PRO GLY SER LEU ILE ASN
SEQRES 22  323  TYR GLY PRO SER GLY ASP GLY SER THR CYS LEU GLY GLY
SEQRES 23  323  ILE GLN SER ASN SER GLY ILE GLY PHE SER ILE PHE GLY
SEQRES 24  323  ASP ILE PHE LEU LYS SER GLN TYR VAL VAL PHE ASP SER
SEQRES 25  323  ASP GLY PRO GLN LEU GLY PHE ALA PRO GLN ALA
```

HELIX	1	H1	ASN	58	GLY	63
HELIX	2	H2	SER	109	ASP	115
HELIX	3	H3	THR	139	LEU	148
HELIX	4	H4	ASP	222	VAL	233
HELIX	5	H5	ASP	239	GLY	243
HELIX	6	H6	GLY	299	GLN	306
SHEET	1	I	GLY	4	ASN	8
SHEET	2	I	GLY	163	GLY	168
SHEET	3	I	PRO	151	LEU	157
SHEET	4	I	SER	305	SER	312
SHEET	5	I	PRO	315	ALA	320
SHEET	6	I	LEU	179	VAL	184
SHEET	1	II	THR	26	LEU	30
SHEET	2	II	THR	19	GLY	24
SHEET	3	II	SER	89	VAL	92
SHEET	4	II	VAL	95	ALA	97
SHEET	1	IIA	THR	9	THR	11
SHEET	2	IIA	GLU	16	ILE	18
SHEET	1	III	GLN	203	PHE	208
SHEET	2	III	VAL	195	GLY	201
SHEET	3	III	ASP	255	ILE	260
SHEET	4	III	TYR	263	VAL	267
SHEET	1	IV	TRP	71	GLY	76
SHEET	2	IV	SER	79	ASN	84
SHEET	3	IV	ALA	104	SER	109
SHEET	1	V	GLN	237	SER	240
SHEET	2	V	GLY	243	PHE	247
SHEET	3	V	THR	282	GLY	285
SHEET	4	V	TYR	274	SER	277
SHEET	1	VI	ASN	29	SER	36
SHEET	2	VI	ASP	119	ALA	125
SHEET	3	VI	ASP	38	PHE	42
SHEET	4	VI	GLN	103	ALA	105
SHEET	1	VII	TRP	191	PHE	193
SHEET	2	VII	GLY	210	THR	217
SHEET	3	VII	SER	296	ILE	301
SHEET	4	VII	LEU	218	LEU	221
SHEET	5	VII	GLY	286	ASN	290

```
HELIX    4   4 ALA    54  VAL    61
HELIX    5   5 SER    64  ILE    75
HELIX    6   6 HIS    80  ILE    87
SHEET    1   I LYS     5  TYR     7
SHEET    2   I PHE    74  HIS  . 80
SHEET    3   I TYR    27  LEU    32
SHEET    4   I THR    21  LEU    25
SHEET    5   I GLY    51  ALA    54
```

```
COMPND   CYTOCHROME $C=3=
HEADER   HEME PROTEIN OF ELECTRON TRANSPORT
SEQRES  1  107  ALA PRO LYS ALA PRO ALA ASP GLY LEU LYS MET ASP LYS
SEQRES  2  107  THR LYS GLN PRO VAL VAL PHE ASN HIS SER THR HIS LYS
SEQRES  3  107  ALA VAL LYS CYS GLY ASP CYS HIS HIS PRO VAL ASN GLY
SEQRES  4  107  LYS GLU ASN TYR GLN LYS CYS ALA THR ALA GLY CYS HIS
SEQRES  5  107  ASP ASN MET ASP LYS LYS ASP LYS SER ALA LYS GLY TYR
SEQRES  6  107  TYR HIS ALA MET HIS ASP LYS GLY THR LYS PHE LYS SER
SEQRES  7  107  CYS VAL GLY CYS HIS LEU GLU THR ALA GLY ALA ASP ALA
SEQRES  8  107  ALA LYS LYS LYS GLU LEU THR GLY CYS LYS GLY SER LYS
SEQRES  9  107  CYS HIS SER
HELIX    1  H1 SER    78  LEU    84
HELIX    2  H2 TYR    65  HIS    70
HELIX    2  H3 LYS    93  THR    98
SHEET    1  S1 LEU     9  MET    11
SHEET    2  S1 VAL    18  PHE    20
```

```
COMPND   CONCANAVALIN A
HEADER   LECTIN (AGGLUTININ)
SEQRES  1  237  ALA ASP THR ILE VAL ALA VAL GLU LEU ASP THR TYR PRO
SEQRES  2  237  ASN THR ASP ILE GLY ASP PRO SER TYR PRO HIS ILE GLY
SEQRES  3  237  ILE ASP ILE LYS SER VAL ARG SER LYS LYS THR ALA LYS
SEQRES  4  237  TRP ASN MET GLN ASP GLY LYS VAL GLY THR ALA HIS ILE
SEQRES  5  237  ILE TYR ASN SER VAL ASP LYS ARG LEU SER ALA VAL VAL
SEQRES  6  237  SER TYR PRO ASN ALA ASP ALA THR SER VAL SER TYR ASP
SEQRES  7  237  VAL ASP LEU ASN ASP VAL LEU PRO GLU TRP VAL ARG VAL
SEQRES  8  237  GLY LEU SER ALA SER THR GLY LEU TYR LYS GLU THR ASN
SEQRES  9  237  THR ILE LEU SER TRP SER PHE THR SER LYS LEU LYS SER
SEQRES 10  237  ASN SER THR HIS GLN THR ASP ALA LEU HIS PHE MET PHE
```

```
SEQRES 11  237  ASN GLN PHE SER LYS ASP GLN LYS ASP LEU ILE LEU GLN
SEQRES 12  237  GLY ASP ALA THR THR GLY THR ASP GLY ASN LEU GLU LEU
SEQRES 13  237  THR ARG VAL SER SER ASN GLY SER PRO GLU GLY SER SER
SEQRES 14  237  VAL GLY ARG ALA LEU PHE TYR ALA PRO VAL HIS ILE TRP
SEQRES 15  237  GLU SER SER ALA THR VAL SER ALA PHE GLU ALA THR PHE
SEQRES 16  237  ALA PHE LEU ILE LYS SER PRO ASP SER HIS PRO ALA ASP
SEQRES 17  237  GLY ILE ALA PHE PHE ILE SER ASN ILE ASP SER SER ILE
SEQRES 18  237  PRO SER GLY SER THR GLY ARG LEU LEU GLY LEU PHE PRO
SEQRES 19  237  ASP ALA ASN
HELIX    1 H1 LEU     81  VAL     84
SHEET    1 BK1  THR    73  VAL    79
SHEET    2 BK1  LYS    59  SER    66
SHEET    3 BK1  GLY    48  SER    56
SHEET    4 BK1  VAL   188  LYS   200
SHEET    5 BK1  SER   108  LYS   116
SHEET    6 BK1  ASP   124  PHE   130
SHEET    1 BK2  THR   103  ILE   106
SHEET    2 BK2  ASN   153  LEU   156
SHEET    3 BK2  THR   147  GLY   149
SHEET    1 FT1  LYS    35  TRP    40
SHEET    2 FT1  PRO    23  LYS    30
SHEET    3 FT1  ILE     4  THR    11
SHEET    4 FT1  ASP   208  SER   215
SHEET    5 FT1  GLU    87  THR    97
SHEET    6 FT1  SER   169  TYR   176
SHEET    7 FT1  ASP   139  GLN   143
SHEET    1 FT2  VAL   179  ILE   181

COMPND  CYTOCHROME P450CAM (CAMPHOR MONOOXYGENASE) (E.C.1.14.15.1)
COMPND  2 WITH BOUND CAMPHOR
HEADER  OXIDOREDUCTASE(OXYGENASE)
SEQRES  1  414  THR THR GLU THR ILE GLN SER ASN ALA ASN LEU ALA PRO
SEQRES  2  414  LEU PRO PRO HIS VAL PRO GLU HIS LEU VAL PHE ASP PHE
SEQRES  3  414  ASP MET TYR ASN PRO SER ASN LEU SER ALA GLY THR GLN
SEQRES  4  414  GLU ALA TRP ALA THR LEU GLN GLU SER ASN VAL PRO ASP
SEQRES  5  414  LEU VAL TRP THR ARG CYS ASN GLY GLY HIS TRP ILE ALA
SEQRES  6  414  THR ARG GLY GLN LEU ILE ARG GLU ALA TYR GLU ASP TYR
SEQRES  7  414  ARG HIS PHE SER SER GLU CYS PRO PHE ILE PRO ARG GLU
```

```
SEQRES  8  414  ALA GLY GLU ALA TYR ASP PHE ILE PRO THR SER MET ASP
SEQRES  9  414  PRO PRO GLU GLN ARG GLN PHE ARG ALA LEU ALA ASN GLN
SEQRES 10  414  VAL VAL GLY MET PRO VAL VAL ASP LYS LEU GLU ASN ARG
SEQRES 11  414  ILE GLN GLU LEU ALA CYS SER LEU ILE GLU SER LEU ARG
SEQRES 12  414  PRO GLN GLY GLN CYS ASN PHE THR GLU ASP TYR ALA GLU
SEQRES 13  414  PRO PHE PRO ILE ARG ILE PHE MET LEU LEU ALA GLY LEU
SEQRES 14  414  PRO GLU GLU ASP ILE PRO HIS LEU LYS TYR LEU THR ASP
SEQRES 15  414  GLN MET THR ARG PRO ASP GLY SER MET THR PHE ALA GLU
SEQRES 16  414  ALA LYS GLU ALA LEU TYR ASP TYR LEU ILE PRO ILE ILE
SEQRES 17  414  GLU GLN ARG ARG GLN ALA PRO GLY THR ASP ALA ILE SER
SEQRES 18  414  ILE VAL ALA ASN GLY GLN VAL ASN GLY ARG PRO ILE THR
SEQRES 19  414  SER ASP GLN ALA LYS ARG MET CYS GLY LEU LEU LEU VAL
SEQRES 20  414  GLY GLY LEU ASP THR VAL VAL ASN PHE LEU SER PHE SER
SEQRES 21  414  MET GLU PHE LEU ALA LYS SER PRO GLU HIS ARG GLN GLU
SEQRES 22  414  LEU ILE GLN ARG PRO GLU ARG ILE PRO ALA ALA CYS GLU
SEQRES 23  414  GLU LEU LEU ARG ARG PHE SER LEU VAL ALA ASP GLY ARG
SEQRES 24  414  ILE LEU THR SER ASP TYR GLU PHE HIS GLY VAL GLN LEU
SEQRES 25  414  LYS LYS GLY ASP GLN ILE LEU LEU PRO GLN MET LEU SER
SEQRES 26  414  GLY LEU ASN GLU ARG GLU ASN ALA CYS PRO MET HIS VAL
SEQRES 27  414  ASP PHE SER ARG GLN LYS VAL SER HIS THR THR PHE GLY
SEQRES 28  414  HIS GLY SER HIS LEU CYS LEU GLY GLN HIS LEU ALA ARG
SEQRES 29  414  ARG GLU ILE ILE VAL THR LEU LYS GLU TRP LEU THR ARG
SEQRES 30  414  ILE PRO ASP PHE SER ILE ALA PRO GLY ALA GLN ILE GLN
SEQRES 31  414  HIS LYS SER GLY ILE VAL SER GLY VAL GLN ALA LEU PRO
SEQRES 32  414  LEU VAL TRP ASP PRO ALA THR THR LYS ALA VAL
HELIX    1     A GLY    37  GLN    46
HELIX    2     B ARG    67  ASP    77
HELIX    3   BND ASP    77  PHE    81
HELIX    4   BPR PRO    89  TYR    96
HELIX    5     C PRO   106  LYS   126
HELIX    6     D LEU   127  GLN   145
HELIX    7     E ASN   149  LEU   169
HELIX    8     F ASP   173  THR   185
HELIX    9     G THR   192  ALA   214
HELIX   10     H ASP   218  ASN   225
HELIX   11     I THR   234  SER   267
HELIX   12     J SER   267  GLN   276
HELIX   13     K ARG   280  PHE   292
```

```
HELIX   14 310 LEU    324  ASN   328
HELIX   15   L GLY    359  ILE   378
SHEET    1  B1  ASP    52  CYS    58
SHEET    2  B1  GLY    60  THR    66
SHEET    1  B5  GLY   146  PHE   150
SHEET    2  B5  ILE   395  VAL   405
SHEET    3  B5  SER   382  SER   397
SHEET    1  B2  GLY   226  VAL   228
SHEET    2  B2  GLY   230  ILE   233
SHEET    1  B3  VAL   295  LEU   301
SHEET    2  B3  GLY   315  MET   323
SHEET    1  B4  TYR   305  HIS   308
SHEET    2  B4  VAL   310  LEU   312

COMPND  CITRATE SYNTHASE (E.C.4.1.3.7) - (CO*A, CITRATE) COMPLEX
HEADER  OXO-ACID-LYASE
SEQRES  1  437  ALA SER SER THR ASN LEU LYS ASP ILE LEU ALA ASP LEU
SEQRES  2  437  ILE PRO LYS GLU GLN ALA ARG ILE LYS THR PHE ARG GLN
SEQRES  3  437  GLN HIS GLY ASN THR ALA VAL GLY GLN ILE THR VAL ASP
SEQRES  4  437  MET MET TYR GLY GLY MET ARG GLY MET LYS GLY LEU VAL
SEQRES  5  437  TYR GLU THR SER VAL LEU ASP PRO ASP GLU GLY ILE ARG
SEQRES  6  437  PHE ARG GLY TYR SER ILE PRO GLU CYS GLN LYS MET LEU
SEQRES  7  437  PRO LYS ALA LYS GLY GLY GLU GLU PRO LEU PRO GLU GLY
SEQRES  8  437  LEU PHE TRP LEU LEU VAL THR GLY GLN ILE PRO THR GLU
SEQRES  9  437  GLU GLN VAL SER TRP LEU SER LYS GLU TRP ALA LYS ARG
SEQRES 10  437  ALA ALA LEU PRO SER HIS VAL VAL THR MET LEU ASP ASN
SEQRES 11  437  PHE PRO THR ASN LEU HIS PRO MET SER GLN LEU SER ALA
SEQRES 12  437  ALA ILE THR ALA LEU ASN SER GLU SER ASN PHE ALA ARG
SEQRES 13  437  ALA TYR ALA GLU GLY ILE HIS ARG THR LYS TYR TRP GLU
SEQRES 14  437  LEU ILE TYR GLU ASP CYS MET ASP LEU ILE ALA LYS LEU
SEQRES 15  437  PRO CYS VAL ALA ALA LYS ILE TYR ARG ASN LEU TYR ARG
SEQRES 16  437  GLU GLY SER SER ILE GLY ALA ILE ASP SER LYS LEU ASP
SEQRES 17  437  TRP SER HIS ASN PHE THR ASN MET LEU GLY TRP THR ASP
SEQRES 18  437  ALA GLN PHE THR GLU LEU MET ARG LEU TYR LEU THR ILE
SEQRES 19  437  HIS SER ASP HIS GLU GLY GLY ASN VAL SER ALA HIS THR
SEQRES 20  437  SER HIS LEU VAL GLY SER ALA LEU SER ASP PRO TYR LEU
SEQRES 21  437  SER PHE ALA ALA ALA MET ASN GLY LEU ALA GLY PRO LEU
SEQRES 22  437  HIS GLY LEU ALA ASN GLN GLU VAL LEU VAL TRP LEU THR
```

```
SEQRES 23  437  GLN LEU GLN LYS GLU VAL GLY LYS ASP VAL SER ASP GLU
SEQRES 24  437  LYS LEU ARG ASP TYR ILE TRP ASN THR LEU ASN SER GLY
SEQRES 25  437  ARG VAL VAL PRO GLY TYR GLY HIS ALA VAL LEU ARG LYS
SEQRES 26  437  THR ASP PRO ARG TYR THR CYS GLN ARG GLU PHE ALA LEU
SEQRES 27  437  LYS HIS LEU PRO HIS ASP PRO MET PHE LYS LEU VAL ALA
SEQRES 28  437  GLN LEU TYR LYS ILE VAL PRO ASN VAL LEU LEU GLU GLN
SEQRES 29  437  GLY LYS ALA LYS ASN PRO TRP PRO ASN VAL ASP ALA HIS
SEQRES 30  437  SER GLY VAL LEU LEU GLN TYR TYR GLY MET THR GLU MET
SEQRES 31  437  ASN TYR TYR THR VAL LEU PHE GLY VAL SER ARG ALA LEU
SEQRES 32  437  GLY VAL LEU ALA GLN LEU ILE TRP SER ARG ALA LEU GLY
SEQRES 33  437  PHE PRO LEU GLU ARG PRO LYS SER MET SER THR ASP GLY
SEQRES 34  437  LEU ILE LYS LEU VAL ASP SER LYS
HELIX    1   A ASN     5  GLY     29
HELIX    2   B THR    37  GLY     43
HELIX    3   C SER    70  LEU     78
HELIX    4   D LEU    88  GLY     99
HELIX    5   E THR   103  ALA    118
HELIX    6   F PRO   121  PHE    131
HELIX    7   G HIS   136  SER    152
HELIX    8   H ASN   153  GLY    161
HELIX    9   I HIS   163  ARG    195
HELIX   10   J ASP   208  GLY    218
HELIX   11   K ASP   221  SER    236
HELIX   12   L ASN   242  LEU    255
HELIX   13   M ASP   257  GLY    271
HELIX   14   N HIS   274  GLU    291
HELIX   15   O SER   297  GLY    312
HELIX   16   P ASP   327  LEU    341
HELIX   17   Q ASP   344  GLY    365
HELIX   18   R ASN   373  GLY    386
HELIX   19   S MET   390  GLY    416
HELIX   20   T SER   426  LEU    433
SHEET    1   A  VAL    57  ASP    59
SHEET    2   A  GLU    62  ARG    65

COMPND  CYTOCHROME $C PEROXIDASE (E.C.1.11.1.5) (FERROCYTOCHROME $C
COMPND  2 (COLON) H2*02 REDUCTASE)
HEADER  OXIDOREDUCTASE (H2O2(A))
```

```
SEQRES  1  294  THR THR PRO LEU VAL HIS VAL ALA SER VAL GLU LYS GLY
SEQRES  2  294  ARG SER TYR GLU ASP PHE GLN LYS VAL TYR ASN ALA ILE
SEQRES  3  294  ALA LEU LYS LEU ARG GLU ASP ASP GLU TYR ASP ASN TYR
SEQRES  4  294  ILE GLY TYR GLY PRO VAL LEU VAL ARG LEU ALA TRP HIS
SEQRES  5  294  THR SER GLY THR TRP ASP LYS HIS ASP ASN THR GLY GLY
SEQRES  6  294  SER TYR GLY GLY THR TYR ARG PHE LYS LYS GLU PHE ASN
SEQRES  7  294  ASP PRO SER ASN ALA GLY LEU GLN ASN GLY PHE LYS PHE
SEQRES  8  294  LEU GLU PRO ILE HIS LYS GLU PHE PRO TRP ILE SER SER
SEQRES  9  294  GLY ASP LEU PHE SER LEU GLY GLY VAL THR ALA VAL GLN
SEQRES 10  294  GLU MET GLN GLY PRO LYS ILE PRO TRP ARG CYS GLY ARG
SEQRES 11  294  VAL ASP THR PRO GLU ASP THR THR PRO ASP ASN GLY ARG
SEQRES 12  294  LEU PRO ASP ALA ASP LYS ASP ALA ASP TYR VAL ARG THR
SEQRES 13  294  PHE PHE GLN ARG LEU ASN MET ASN ASP ARG GLU VAL VAL
SEQRES 14  294  ALA LEU MET GLY ALA HIS ALA LEU GLY LYS THR HIS LEU
SEQRES 15  294  LYS ASN SER GLY TYR GLU GLY PRO TRP GLY ALA ALA ASN
SEQRES 16  294  ASN VAL PHE THR ASN GLU PHE TYR LEU ASN LEU LEU ASN
SEQRES 17  294  GLU ASP TRP LYS LEU GLU LYS ASN ASP ALA ASN ASN GLU
SEQRES 18  294  GLN TRP ASP SER LYS SER GLY TYR MET MET LEU PRO THR
SEQRES 19  294  ASP TYR SER LEU ILE GLN ASP PRO LYS TYR LEU SER ILE
SEQRES 20  294  VAL LYS GLU TYR ALA ASN ASP GLN ASP LYS PHE PHE LYS
SEQRES 21  294  ASP PHE SER LYS ALA PHE GLU LYS LEU LEU GLU ASP GLY
SEQRES 22  294  ILE THR PHE PRO LYS ASP ALA PRO SER PRO PHE ILE PHE
SEQRES 23  294  LYS THR LEU GLU GLU GLN GLY LEU
HELIX    1   A SER     15  ASP     33
HELIX    2   B TYR     42  SER     54
HELIX    3  B1 PHE     73  ASN     78
HELIX    4   C GLY     84  PHE     99
HELIX    5   D SER    103  MET    119
HELIX    6   E ASP    150  PHE    158
HELIX    7   F ASN    164  LEU    177
HELIX    8  F1 HIS    181  GLY    186
HELIX    9   G GLU    201  GLU    209
HELIX   10   H LEU    232  GLN    240
HELIX   11   I ASP    241  ASN    253
HELIX   12   J GLN    255  ASP    272
HELIX   13  J1 THR    288  GLY    293
```

```
COMPND   ERABUTOXIN $B
HEADER   TOXIN
SEQRES  1  62   ARG ILE CYS PHE ASN HIS GLN SER SER GLN PRO GLN THR
SEQRES  2  62   THR LYS THR CYS SER PRO GLY GLU SER SER CYS TYR HIS
SEQRES  3  62   LYS GLN TRP SER ASP PHE ARG GLY THR ILE ILE GLU ARG
SEQRES  4  62   GLY CYS GLY CYS PRO THR VAL LYS PRO GLY ILE LYS LEU
SEQRES  5  62   SER CYS CYS GLU SER GLU VAL CYS ASN ASN
SHEET   1  AB   ARG     1  ASN     5
SHEET   2  AB   GLN    12  CYS    17

COMPND   HEMOGLOBIN V (CYANO,MET)
HEADER   OXYGEN TRANSPORT
SEQRES  1  149  PRO ILE VAL ASP THR GLY SER VAL ALA PRO LEU SER ALA
SEQRES  2  149  ALA GLU LYS THR LYS ILE ARG SER ALA TRP ALA PRO VAL
SEQRES  3  149  TYR SER THR TYR GLU THR SER GLY VAL ASP ILE LEU VAL
SEQRES  4  149  LYS PHE PHE THR SER THR PRO ALA ALA GLN GLU PHE PHE
SEQRES  5  149  PRO LYS PHE LYS GLY LEU THR THR ALA ASP GLU LEU LYS
SEQRES  6  149  LYS SER ALA ASP VAL ARG TRP HIS ALA GLU ARG ILE ILE
SEQRES  7  149  ASN ALA VAL ASP ASP ALA VAL ALA SER MET ASP ASP THR
SEQRES  8  149  GLU LYS MET SER MET LYS LEU ARG ASN LEU SER GLY LYS
SEQRES  9  149  HIS ALA LYS SER PHE GLN VAL ASP PRO GLU TYR PHE LYS
SEQRES 10  149  VAL LEU ALA ALA VAL ILE ALA ASP THR VAL ALA ALA GLY
SEQRES 11  149  ASP ALA GLY PHE GLU LYS LEU MET SER MET ILE CYS ILE
SEQRES 12  149  LEU LEU ARG SER ALA TYR
HELIX   1   A ALA     13  SER     28
HELIX   2   B TYR     30  SER     44
HELIX   3   C PRO     46  PHE     51
HELIX   4   D ALA     61  LYS     66
HELIX   5   E ALA     68  ALA     86
HELIX   6   F THR     91  PHE    109
HELIX   7   G PRO    113  VAL    127
HELIX   8   H ALA    132  ALA    148

COMPND   LYSOZYME (E.C.3.2.1.17)
HEADER   HYDROLASE (O-GLYCOSYL)
SEQRES  1  164  MET ASN ILE PHE GLU MET LEU ARG ILE ASP GLU GLY LEU
SEQRES  2  164  ARG LEU LYS ILE TYR LYS ASP THR GLU GLY TYR TYR THR
```

```
SEQRES  3  164  ILE GLY ILE GLY HIS LEU LEU THR LYS SER PRO SER LEU
SEQRES  4  164  ASN ALA ALA LYS SER GLU LEU ASP LYS ALA ILE GLY ARG
SEQRES  5  164  ASN CYS ASN GLY VAL ILE THR LYS ASP GLU ALA GLU LYS
SEQRES  6  164  LEU PHE ASN GLN ASP VAL ASP ALA ALA VAL ARG GLY ILE
SEQRES  7  164  LEU ARG ASN ALA LYS LEU LYS PRO VAL TYR ASP SER LEU
SEQRES  8  164  ASP ALA VAL ARG ARG CYS ALA LEU ILE ASN MET VAL PHE
SEQRES  9  164  GLN MET GLY GLU THR GLY VAL ALA GLY PHE THR ASN SER
SEQRES 10  164  LEU ARG MET LEU GLN GLN LYS ARG TRP ASP GLU ALA ALA
SEQRES 11  164  VAL ASN LEU ALA LYS SER ARG TRP TYR ASN GLN THR PRO
SEQRES 12  164  ASN ARG ALA LYS ARG VAL ILE THR THR PHE ARG THR GLY
SEQRES 13  164  THR TRP ASP ALA TYR LYS ASN LEU
HELIX    1  H1 ILE      3  GLU     11
HELIX    2  H2 LEU     39  ILE     50
HELIX    3  H3 LYS     60  ARG     80
HELIX    4  H4 ALA     82  SER     90
HELIX    5  H5 ALA     93  MET    106
HELIX    6  H6 GLU    108  GLY    113
HELIX    7  H7 THR    115  GLN    123
HELIX    8  H8 TRP    126  ALA    134
HELIX    9  H9 ARG    137  GLN    141
HELIX   10 H10 PRO    143  THR    155
SHEET    1  S1  GLY    56  ILE    58
SHEET    2  S1  ARG    14  ASP    20
SHEET    3  S1  TYR    24  ILE    27
SHEET    4  S1  HIS    31  THR    34

COMPND  MYOHEMERYTHRIN
HEADER  OXYGEN BINDING
SEQRES  1  118  GLY TRP GLU ILE PRO GLU PRO TYR VAL TRP ASP GLU SER
SEQRES  2  118  PHE ARG VAL PHE TYR GLU GLN LEU ASP GLU GLU HIS LYS
SEQRES  3  118  LYS ILE PHE LYS GLY ILE PHE ASP CYS ILE ARG ASP ASN
SEQRES  4  118  SER ALA PRO ASN LEU ALA THR LEU VAL LYS VAL THR THR
SEQRES  5  118  ASN HIS PHE THR HIS GLU GLU ALA MET MET ASP ALA ALA
SEQRES  6  118  LYS TYR SER GLU VAL VAL PRO HIS LYS LYS MET HIS LYS
SEQRES  7  118  ASP PHE LEU GLU LYS ILE GLY GLY LEU SER ALA PRO VAL
SEQRES  8  118  ASP ALA LYS ASN VAL ASP TYR CYS LYS GLU TRP LEU VAL
SEQRES  9  118  ASN HIS ILE LYS GLY THR ASP PHE LYS TYR LYS GLY LYS
SEQRES 10  118  LEU
```

```
HELIX    1    A GLU      19   ARG      37
HELIX    2    B ALA      41   ALA      64
HELIX    3    C VAL      71   ILE      84
HELIX    4    D ALA      93   LYS     108
HELIX    5 310 ASP      111   LYS     113

COMPND   OVOMUCOID THIRD DOMAIN
HEADER   PROTEINASE INHIBITOR (KAZAL)
SEQRES   1   56   LEU ALA ALA VAL SER VAL ASP CYS SER GLU TYR PRO LYS
SEQRES   2   56   PRO ALA CYS THR MET GLU TYR ARG PRO LEU CYS GLY SER
SEQRES   3   56   ASP ASN LYS THR TYR GLY ASN LYS CYS ASN PHE CYS ASN
SEQRES   4   56   ALA VAL VAL GLU SER ASN GLY THR LEU THR LEU SER HIS
SEQRES   5   56   PHE GLY LYS CYS
HELIX    1   HA ASN      33   SER      44
SHEET    1   SA  LYS     29   GLY      32
SHEET    2   SA  ARG     21   SER      26
SHEET    3   SA  THR     49   GLY      54

COMPND   PREALBUMIN (HUMAN PLASMA)
HEADER   TRANSPORT (THYROXINE,RETINOL) IN SERUM
SEQRES   1   127   GLY PRO THR GLY THR GLY GLU SER LYS CYS PRO LEU MET
SEQRES   2   127   VAL LYS VAL LEU ASP ALA VAL ARG GLY SER PRO ALA ILE
SEQRES   3   127   ASN VAL ALA VAL HIS VAL PHE ARG LYS ALA ALA ASP ASP
SEQRES   4   127   THR TRP GLU PRO PHE ALA SER GLY LYS THR SER GLU SER
SEQRES   5   127   GLY GLU LEU HIS GLY LEU THR THR GLU GLU GLN PHE VAL
SEQRES   6   127   GLU GLY ILE TYR LYS VAL GLU ILE ASP THR LYS SER TYR
SEQRES   7   127   TRP LYS ALA LEU GLY ILE SER PRO PHE HIS GLU HIS ALA
SEQRES   8   127   GLU VAL VAL PHE THR ALA ASN ASP SER GLY PRO ARG ARG
SEQRES   9   127   TYR THR ILE ALA ALA LEU LEU SER PRO TYR SER TYR SER
SEQRES  10   127   THR THR ALA VAL VAL THR ASN PRO LYS GLU
HELIX    1   AA THR      75   GLY      83
SHEET    1 INT   GLY     53   LEU      55
SHEET    2 INT   PRO     11   ALA      19
SHEET    3 INT   TYR    105   SER     112
SHEET    4 INT   TYR    114   VAL     121
SHEET    1 X1A   GLU     42   GLU      42
SHEET    2 X1A   VAL     28   LYS      35
SHEET    3 X1A   GLY     67   ASP      74
```

```
SHEET    4 X1A  ALA   91  ALA   97
SHEET    1 X2A  ALA   45  THR   49

COMPND   KALLIKREIN A (E.C.3.4.21.8)
HEADER   SERINE PROTEINASE
SEQRES   1 A    80   ILE ILE GLY GLY ARG GLU CYS GLU LYS ASN SER HIS PRO
SEQRES   2 A    80   TRP GLN VAL ALA ILE TYR HIS TYR SER SER PHE GLN CYS
SEQRES   3 A    80   GLY GLY VAL LEU VAL ASN PRO LYS TRP VAL LEU THR ALA
SEQRES   4 A    80   ALA HIS CYS LYS ASN ASP ASN TYR GLU VAL TRP LEU GLY
SEQRES   5 A    80   ARG HIS ASN LEU PHE GLU ASN GLU ASN THR ALA GLN PHE
SEQRES   6 A    80   PHE GLY VAL THR ALA ASP PHE PRO HIS PRO GLY PHE ASN
SEQRES   7 A    80   LEU SER
SEQRES   1 B   152   ALA ASP GLY LYS ASP TYR SER HIS ASP LEU MET LEU LEU
SEQRES   2 B   152   ARG LEU GLN SER PRO ALA LYS ILE THR ASP ALA VAL LYS
SEQRES   3 B   152   VAL LEU GLU LEU PRO THR GLN GLU PRO GLU LEU GLY SER
SEQRES   4 B   152   THR CYS GLU ALA SER GLY TRP GLY SER ILE GLU PRO GLY
SEQRES   5 B   152   PRO ASP ASP PHE GLU PHE PRO ASP GLU ILE GLN CYS VAL
SEQRES   6 B   152   GLN LEU THR LEU LEU GLN ASN THR PHE CYS ALA ASP ALA
SEQRES   7 B   152   HIS PRO ASP LYS VAL THR GLU SER MET LEU CYS ALA GLY
SEQRES   8 B   152   TYR LEU PRO GLY GLY LYS ASP THR CYS MET GLY ASP SER
SEQRES   9 B   152   GLY GLY PRO LEU ILE CYS ASN GLY MET TRP GLN GLY ILE
SEQRES  10 B   152   THR SER TRP GLY HIS THR PRO CYS GLY SER ALA ASN LYS
SEQRES  11 B   152   PRO SER ILE TYR THR LYS LEU ILE PHE TYR LEU ASP TRP
SEQRES  12 B   152   ILE ASP ASP THR ILE THR GLU ASN PRO

COMPND   BENCE-*JONES PROTEIN (LAMBDA, VARIABLE DOMAIN)
HEADER   IMMUNOGLOBULIN
SEQRES   1   114   GLU SER VAL LEU THR GLN PRO PRO SER ALA SER GLY THR
SEQRES   2   114   PRO GLY GLN ARG VAL THR ILE SER CYS THR GLY SER ALA
SEQRES   3   114   THR ASP ILE GLY SER ASN SER VAL ILE TRP TYR GLN GLN
SEQRES   4   114   VAL PRO GLY LYS ALA PRO LYS LEU LEU ILE TYR TYR ASN
SEQRES   5   114   ASP LEU LEU PRO SER GLY VAL SER ASP ARG PHE SER ALA
SEQRES   6   114   SER LYS SER GLY THR SER ALA SER LEU ALA ILE SER GLY
SEQRES   7   114   LEU GLU SER GLU ASP GLU ALA ASP TYR TYR CYS ALA ALA
SEQRES   8   114   TRP ASN ASP SER LEU ASP GLU PRO GLY PHE GLY GLY GLY
SEQRES   9   114   THR LYS LEU THR VAL LEU GLY GLN PRO LYS
HELIX    1   A SER    25  ASN    32
SHEET    1   A  GLY   15  GLY    24
```

SHEET	2	A	THR	70	LEU	79
SHEET	3	A	ARG	62	SER	68
SHEET	1	B	LYS	46	ILE	49
SHEET	2	B	ILE	35	GLN	39
SHEET	3	B	ALA	85	ALA	91
SHEET	4	B	GLY	100	LEU	107

COMPND STAPHYLOCOCCAL NUCLEASE (E.C.3.1.4.7) COMPLEX WITH
COMPND 2 2(PRIME)-DEOXY-3(PRIME)-5(PRIME)-DIPHOSPHOTHYMIDINE
HEADER HYDROLASE (PHOSPHORIC DIESTER)

SEQRES	1	149	ALA	THR	SER	THR	LYS	LYS	LEU	HIS	LYS	GLU	PRO	ALA	THR
SEQRES	2	149	LEU	ILE	LYS	ALA	ILE	ASP	GLY	ASP	THR	VAL	LYS	LEU	MET
SEQRES	3	149	TYR	LYS	GLY	GLN	PRO	MET	THR	PHE	ARG	LEU	LEU	LEU	VAL
SEQRES	4	149	ASP	THR	PRO	GLU	THR	LYS	HIS	PRO	LYS	LYS	GLY	VAL	GLU
SEQRES	5	149	LYS	TYR	GLY	PRO	GLU	ALA	SER	ALA	PHE	THR	LYS	LYS	MET
SEQRES	6	149	VAL	GLU	ASN	ALA	LYS	LYS	ILE	GLU	VAL	GLU	PHE	ASN	LYS
SEQRES	7	149	GLY	GLN	ARG	THR	ASP	LYS	TYR	GLY	ARG	GLY	LEU	ALA	TYR
SEQRES	8	149	ILE	TYR	ALA	ASP	GLY	LYS	MET	VAL	ASN	GLU	ALA	LEU	VAL
SEQRES	9	149	ARG	GLN	GLY	LEU	ALA	LYS	VAL	ALA	TYR	VAL	TYR	LYS	PRO
SEQRES	10	149	ASN	ASN	THR	HIS	GLU	GLN	HIS	LEU	ARG	LYS	SER	GLU	ALA
SEQRES	11	149	GLN	ALA	LYS	LYS	GLU	LYS	LEU	ASN	ILE	TRP	SER	GLU	ASN
SEQRES	12	149	ASP	ALA	ASP	SER	GLY	GLN							

HELIX	1	H1	GLY	55	GLU .	67
HELIX	2	H2	VAL	99	GLN	106
HELIX	3	H3	GLU	122	LYS	134
SHEET	1	B1	ALA	12	ASP	19
SHEET	2	B1	ASP	21	TYR	27
SHEET	3	B1	GLN	30	LEU	36
SHEET	1	B2	LYS	71	GLU	75
SHEET	2	B2	TYR	91	ASP	95

COMPND $TRP REPRESSOR (ORTHORHOMBIC FORM)
HEADER DNA BINDING REGULATORY PROTEIN

SEQRES	1	107	ALA	GLN	GLN	SER	PRO	TYR	SER	ALA	ALA	MET	ALA	GLU	GLN
SEQRES	2	107	ARG	HIS	GLN	GLU	TRP	LEU	ARG	PHE	VAL	ASP	LEU	LEU	LYS
SEQRES	3	107	ASN	ALA	TYR	GLN	ASN	ASP	LEU	HIS	LEU	PRO	LEU	LEU	ASN
SEQRES	4	107	LEU	MET	LEU	THR	PRO	ASP	GLU	ARG	GLU	ALA	LEU	GLY	THR
SEQRES	5	107	ARG	VAL	ARG	ILE	VAL	GLU	GLU	LEU	LEU	ARG	GLY	GLU	MET

```
SEQRES  6  107  SER GLN ARG GLU LEU LYS ASN GLU LEU GLY ALA GLY ILE
SEQRES  7  107  ALA THR ILE THR ARG GLY SER ASN SER LEU LYS ALA ALA
SEQRES  8  107  PRO VAL GLU LEU ARG GLN TRP LEU GLU GLU VAL LEU LEU
SEQRES  9  107  LYS SER ASP
HELIX   1   A ALA    12  GLN    31
HELIX   2   B`HIS    35  MET    42
HELIX   3   C ASP    46  ARG    63
HELIX   4   D GLN    68  GLU    74
HELIX   5   E ILE    79  ALA    91
HELIX   6   F VAL    94  LEU   105

COMPND  CYTOCHROME $C=551= (OXIDIZED)
HEADER  ELECTRON TRANSPORT
SEQRES  1  82  GLU ASP PRO GLU VAL LEU PHE LYS ASN LYS GLY CYS VAL
SEQRES  2  82  ALA CYS HIS ALA ILE ASP THR LYS MET VAL GLY PRO ALA
SEQRES  3  82  TYR LYS ASP VAL ALA ALA LYS PHE ALA GLY GLN ALA GLY
SEQRES  4  82  ALA GLU ALA GLU LEU ALA GLN ARG ILE LYS ASN GLY SER
SEQRES  5  82  GLN GLY VAL TRP GLY PRO ILE PRO MET PRO PRO ASN ALA
SEQRES  6  82  VAL SER ASP ASP GLU ALA GLN THR LEU ALA LYS TRP VAL
SEQRES  7  82  LEU SER GLN LYS
HELIX   1   N PRO     3  ASN     9
HELIX   2 310 GLY    11  CYS    15
HELIX   3  30 TYR    27  LYS    33
HELIX   4  40 ALA    40  LYS    49
HELIX   5   C ASP    68  SER    80

COMPND  CYTOCHROME $C=2= (REDUCED)
HEADER  ELECTRON TRANSPORT PROTEIN (CYTOCHROME)
SEQRES  1  112  GLU GLY ASP ALA ALA ALA GLY GLU LYS VAL SER LYS LYS
SEQRES  2  112  CYS LEU ALA CYS HIS THR PHE ASP GLN GLY GLY ALA ASN
SEQRES  3  112  LYS VAL GLY PRO ASN LEU PHE GLY VAL PHE GLU ASN THR
SEQRES  4  112  ALA ALA HIS LYS ASP ASN TYR ALA TYR SER GLU SER TYR
SEQRES  5  112  THR GLU MET LYS ALA LYS GLY LEU THR TRP THR GLU ALA
SEQRES  6  112  ASN LEU ALA ALA TYR VAL LYS ASN PRO LYS ALA PHE VAL
SEQRES  7  112  LEU GLU LYS SER GLY ASP PRO LYS ALA LYS SER LYS MET
SEQRES  8  112  THR PHE LYS LEU THR LYS ASP ASP GLU ILE GLU ASN VAL
SEQRES  9  112  ILE ALA TYR LEU LYS THR LEU LYS
HELIX   1   A ASP     3  SER    11
```

```
HELIX    2 310 SER     11  CYS     14
HELIX    3  A1 CYS     14  HIS     18
HELIX    4   B GLU     50  LYS     58
HELIX    5   C THR     63  GLY     83
HELIX    6   D LYS     97  LYS    109

COMPND  DIHYDROFOLATE REDUCTASE (E.C.1.5.1.3) COMPLEX WITH NADPH AND
COMPND  2 METHOTREXATE
HEADER  OXIDO-REDUCTASE
SEQRES  1  162  THR ALA PHE LEU TRP ALA GLN ASN ARG ASN GLY LEU ILE
SEQRES  2  162  GLY LYS ASP GLY HIS LEU PRO TRP HIS LEU PRO ASP ASP
SEQRES  3  162  LEU HIS TYR PHE ARG ALA GLN THR VAL GLY LYS ILE MET
SEQRES  4  162  VAL VAL GLY ARG ARG THR TYR GLU SER PHE PRO LYS ARG
SEQRES  5  162  PRO LEU PRO GLU ARG THR ASN VAL VAL LEU THR HIS GLN
SEQRES  6  162  GLU ASP TYR GLN ALA GLN GLY ALA VAL VAL VAL HIS ASP
SEQRES  7  162  VAL ALA ALA VAL PHE ALA TYR ALA LYS GLN HIS LEU ASP
SEQRES  8  162  GLN GLU LEU VAL ILE ALA GLY GLY ALA GLN ILE PHE THR
SEQRES  9  162  ALA PHE LYS ASP ASP VAL ASP THR LEU LEU VAL THR ARG
SEQRES 10  162  LEU ALA GLY SER PHE GLU GLY ASP THR LYS MET ILE PRO
SEQRES 11  162  LEU ASN TRP ASP ASP PHE THR LYS VAL SER SER ARG THR
SEQRES 12  162  VAL GLU ASP THR ASN PRO ALA LEU THR HIS THR TYR GLU
SEQRES 13  162  VAL TRP GLN LYS LYS ALA
HELIX    1  HB LEU     23  THR     34
HELIX    2  HC GLY     42  PHE     49
HELIX    3  HE ASP     78  HIS     89
HELIX    4  HF GLY     99  LYS    107
SHEET    1  S1  VAL    74  VAL     76
SHEET    2  S1  GLU    56  THR     63
SHEET    3  S1  GLY    36  GLY     42
SHEET    4  S1  GLU    93  GLY     98
SHEET    5  S1  THR     1  GLN      7
SHEET    6  S1  ASP   111  ALA    119
SHEET    7  S1  THR   152  LYS    161
SHEET    8  S1  ASP   135  VAL    144

COMPND  GLUTATHIONE REDUCTASE (E.C.1.6.4.2), OXIDIZED FORM (E)
HEADER  OXIDOREDUCTASE (FLAVOENZYME)
SEQRES  1  478  ALA CYS ARG GLN GLU PRO GLN PRO GLN GLY PRO PRO PRO
```

```
SEQRES  2  478  ALA ALA GLY ALA VAL ALA SER TYR ASP TYR LEU VAL ILE
SEQRES  3  478  GLY GLY GLY SER GLY GLY LEU ALA SER ALA ARG ARG ALA
SEQRES  4  478  ALA GLU LEU GLY ALA ARG ALA ALA VAL VAL GLU SER HIS
SEQRES  5  478  LYS LEU GLY GLY THR CYS VAL ASN VAL GLY CYS VAL PRO
SEQRES  6  478  LYS LYS VAL MET TRP ASN THR ALA VAL HIS SER GLU PHE
SEQRES  7  478  MET HIS ASP HIS ALA ASP TYR GLY PHE PRO SER CYS GLU
SEQRES  8  478  GLY LYS PHE ASN TRP ARG VAL ILE LYS GLU LYS ARG ASP
SEQRES  9  478  ALA TYR VAL SER ARG LEU ASN ALA ILE TYR GLN ASN ASN
SEQRES 10  478  LEU THR LYS SER HIS ILE GLU ILE ILE ARG GLY HIS ALA
SEQRES 11  478  ALA PHE THR SER ASP PRO LYS PRO THR ILE GLU VAL SER
SEQRES 12  478  GLY LYS LYS TYR THR ALA PRO HIS ILE LEU ILE ALA THR
SEQRES 13  478  GLY GLY MET PRO SER THR PRO HIS GLU SER GLN ILE PRO
SEQRES 14  478  GLY ALA SER LEU GLY ILE THR SER ASP GLY PHE PHE GLN
SEQRES 15  478  LEU GLU GLU LEU PRO GLY ARG SER VAL ILE VAL GLY ALA
SEQRES 16  478  GLY TYR ILE ALA VAL GLU MET ALA GLY ILE LEU SER ALA
SEQRES 17  478  LEU GLY SER LYS THR SER LEU MET ILE ARG HIS ASP LYS
SEQRES 18  478  VAL LEU ARG SER PHE ASP SER MET ILE SER THR ASN CYS
SEQRES 19  478  THR GLU GLU LEU GLU ASN ALA GLY VAL GLU VAL LEU LYS
SEQRES 20  478  PHE SER GLN VAL LYS GLU VAL LYS LYS THR LEU SER GLY
SEQRES 21  478  LEU GLU VAL SER MET VAL THR ALA VAL PRO GLY ARG LEU
SEQRES 22  478  PRO VAL MET THR MET ILE PRO ASP VAL ASP CYS LEU LEU
SEQRES 23  478  TRP ALA ILE GLY ARG VAL PRO ASN THR LYS ASP LEU SER
SEQRES 24  478  LEU ASN LYS LEU GLY ILE GLN THR ASP ASP LYS GLY HIS
SEQRES 25  478  ILE ILE VAL ASP GLU PHE GLN ASN THR ASN VAL LYS GLY
SEQRES 26  478  ILE TYR ALA VAL GLY ASP VAL CYS GLY LYS ALA LEU LEU
SEQRES 27  478  THR PRO VAL ALA ILE ALA ALA GLY ARG LYS LEU ALA HIS
SEQRES 28  478  ARG LEU PHE GLU TYR LYS GLU ASP SER LYS LEU ASP TYR
SEQRES 29  478  ASN ASN ILE PRO THR VAL VAL PHE SER HIS PRO PRO ILE
SEQRES 30  478  GLY THR VAL GLY LEU THR GLU ASP GLU ALA ILE HIS LYS
SEQRES 31  478  TYR GLY ILE GLU ASN VAL LYS THR TYR SER THR SER PHE
SEQRES 32  478  THR PRO MET TYR HIS ALA VAL THR LYS ARG LYS THR LYS
SEQRES 33  478  CYS VAL MET LYS MET VAL CYS ALA ASN LYS GLU GLU LYS
SEQRES 34  478  VAL VAL GLY ILE HIS MET GLN GLY LEU GLY CYS ASP GLU
SEQRES 35  478  MET LEU GLN GLY PHE ALA VAL ALA VAL LYS MET GLY ALA
SEQRES 36  478  THR LYS ALA ASP PHE ASP ASN THR VAL ALA ILE HIS PRO
SEQRES 37  478  THR SER SER GLU GLU LEU VAL THR LEU ARG
HELIX    1  H1 GLY     29  GLY     43
HELIX    2  H2 GLY     56  GLY     86
```

HELIX	3	H3	TRP	96	HIS	122
HELIX	4	H4	GLY	170	GLY	174
HELIX	5	H5	SER	177	LEU	183
HELIX	6	H6	GLY	196	LEU	209
HELIX	7	H7	ASP	227	GLY	242
HELIX	8	H8	SER	299	GLY	304
HELIX	9	H9	GLY	330	GLY	334
HELIX	10	H10	LEU	338	PHE	354
HELIX	11	H11	THR	383	GLY	392
HELIX	12	H12	PRO	405	ALA	409
HELIX	13	H13	LEU	444	MET	453
HELIX	14	H14	THR	456	ASN	462
HELIX	15	H15	SER	470	THR	476
SHEET	1	A	GLU	124	GLY	128
SHEET	2	A	ARG	45	GLU	50
SHEET	3	A	ASP	22	GLY	27
SHEET	4	A	HIS	151	ALA	155
SHEET	5	A	GLY	325	VAL	329
SHEET	6	A	GLN	319	VAL	323
SHEET	7	A	GLY	311	ILE	314
SHEET	8	A	GLN	306	ASP	308
SHEET	1	B	ALA	19	TYR	21
SHEET	2	B	LYS	145	THR	148
SHEET	3	B	THR	139	VAL	142
SHEET	4	B	ALA	131	THR	133
SHEET	1	C	ILE	175	ILE	175
SHEET	2	C	ASP	283	ALA	288
SHEET	3	C	GLY	188	GLY	194
SHEET	4	C	LYS	212	ILE	217
SHEET	5	C	GLU	244	SER	249
SHEET	1	D	PRO	169	PRO	169
SHEET	2	D	PHE	248	THR	257
SHEET	3	D	GLY	260	ALA	268
SHEET	4	D	VAL	275	VAL	282
SHEET	1	E	THR	369	VAL	371
SHEET	2	E	PRO	376	LEU	382
SHEET	3	E	GLU	428	GLY	437
SHEET	4	E	CYS	417	ASN	425

SHEET	5	E	ASN	395	PHE	403
SHEET	1	F	GLY	157	SER	161
SHEET	2	F	GLY	290	ASN	294

COMPND RAT MAST CELL PROTEASE /II$ (/RMCPII$)

HEADER SERINE PROTEINASE

SEQRES	1	224	ILE ILE GLY GLY VAL GLU SER ILE PRO HIS SER ARG PRO
SEQRES	2	224	TYR MET ALA HIS LEU ASP ILE VAL THR GLU LYS GLY LEU
SEQRES	3	224	ARG VAL ILE CYS GLY GLY PHE LEU ILE SER ARG GLN PHE
SEQRES	4	224	VAL LEU THR ALA ALA HIS CYS LYS GLY ARG GLU ILE THR
SEQRES	5	224	VAL ILE LEU GLY ALA HIS ASP VAL ARG LYS ALA GLU SER
SEQRES	6	224	THR GLN GLN LYS ILE LYS VAL GLU LYS GLN ILE ILE HIS
SEQRES	7	224	GLU SER TYR ASN SER ALA PRO ARG LEU HIS ASP ILE MET
SEQRES	8	224	LEU LEU LYS LEU GLU LYS LYS VAL GLU LEU THR PRO ALA
SEQRES	9	224	VAL ASN VAL VAL PRO LEU PRO SER PRO SER ASP PHE ILE
SEQRES	10	224	HIS PRO GLY ALA MET CYS TRP ALA ALA GLY TRP GLY LYS
SEQRES	11	224	THR GLY VAL ARG ASP PRO THR SER TYR THR LEU ARG GLU
SEQRES	12	224	VAL GLU LEU ARG ILE MET ASP GLU LYS ALA CYS VAL ASP
SEQRES	13	224	TYR GLY TYR TYR GLU TYR LYS PHE GLN VAL CYS VAL GLY
SEQRES	14	224	SER PRO THR THR LEU ARG ALA ALA PHE MET GLY ASP SER
SEQRES	15	224	GLY GLY PRO LEU LEU CYS ALA GLY VAL ALA HIS GLY ILE
SEQRES	16	224	VAL SER TYR GLY HIS PRO ASP ALA LYS PRO PRO ALA ILE
SEQRES	17	224	PHE THR ARG VAL SER THR TYR VAL PRO TRP ILE ASN ALA
SEQRES	18	224	VAL VAL ASN

COMPND RUBREDOXIN

HEADER ELECTRON TRANSPORT(NON-HEME FE PROTEIN)

SEQRES	1	52	MET LYS LYS TYR VAL CYS THR VAL CYS GLY TYR GLU TYR
SEQRES	2	52	ASP PRO ALA GLU GLY ASP PRO THR ASN GLY VAL LYS PRO
SEQRES	3	52	GLY THR SER PHE ASP ASP LEU PRO ALA ASP TRP VAL CYS
SEQRES	4	52	PRO VAL CYS GLY ALA PRO LYS SER GLU PHE GLU ALA ALA

COMPND WHEAT GERM AGGLUTININ (ISOLECTIN 2)

HEADER LECTIN (AGGLUTININ)

SEQRES	1	170	ARG CYS GLY GLU GLN GLY SER ASN MET GLU CYS PRO ASN
SEQRES	2	170	ASN LEU CYS CYS SER GLN TYR GLY TYR CYS GLY MET GLY
SEQRES	3	170	GLY ASP TYR CYS GLY LYS GLY CYS GLN ASP GLY ALA CYS
SEQRES	4	170	TRP THR SER LYS ARG CYS GLY SER GLN ALA GLY GLY ALA

```
SEQRES  5   170  THR CYS PRO ASN ASN HIS CYS CYS SER GLN TYR GLY HIS
SEQRES  6   170  CYS GLY PHE GLY ALA GLU TYR CYS GLY ALA GLY CYS GLN
SEQRES  7   170  GLY GLY PRO CYS ARG ALA ASP ILE LYS CYS GLY SER GLN
SEQRES  8   170  SER GLY GLY LYS LEU CYS PRO ASN ASN LEU CYS CYS SER
SEQRES  9   170  GLN TRP GLY SER CYS GLY LEU GLY SER GLU PHE CYS GLY
SEQRES 10   170  GLY GLY CYS GLN SER GLY ALA CYS SER THR ASP LYS PRO
SEQRES 11   170  CYS GLY GLY ASP ALA GLY GLY ARG VAL CYS THR ASN ASN
SEQRES 12   170  TYR CYS CYS SER ALA GLY GLY SER CYS GLY ILE GLY PRO
SEQRES 13   170  GLY TYR CYS GLY ALA GLY CYS GLN SER GLY GLY CYS ASP
SEQRES 14   170  GLY
HELIX    1  AA GLY    27  GLY    31
HELIX    2  BA GLY    69  GLY    74
HELIX    3  CA GLY   112  GLY   117
HELIX    4  DA GLY   155  GLY   160

COMPND  CARBOXYPEPTIDASE A=ALPHA= (COX) (E.C.3.4.17.1) COMPLEX WITH
COMPND  2 POTATO CARBOXYPEPTIDASE A INHIBITOR
HEADER  HYDROLASE (C-TERMINAL PEPTIDASE)
SEQRES  1  307  ALA ARG SER THR ASN THR PHE ASN TYR ALA THR TYR HIS
SEQRES  2  307  THR LEU ASP GLU ILE TYR ASP PHE MET ASP LEU LEU VAL
SEQRES  3  307  ALA GLN HIS PRO GLU LEU VAL SER LYS LEU GLN ILE GLY
SEQRES  4  307  ARG SER TYR GLU GLY ARG PRO ILE TYR VAL LEU LYS PHE
SEQRES  5  307  SER THR GLY GLY SER ASN ARG PRO ALA ILE TRP ILE ASP
SEQRES  6  307  LEU GLY ILE HIS SER ARG GLN TRP ILE THR GLN ALA THR
SEQRES  7  307  GLY VAL TRP PHE ALA LYS LYS PHE THR GLU ASN TYR GLY
SEQRES  8  307  GLN ASN PRO SER PHE THR ALA ILE LEU ASP SER MET ASP
SEQRES  9  307  ILE PHE LEU GLU ILE VAL THR ASN PRO ASN GLY PHE ALA
SEQRES 10  307  PHE THR HIS SER GLU ASN ARG LEU TRP ARG LYS THR ARG
SEQRES 11  307  SER VAL THR SER SER SER LEU CYS VAL GLY VAL ASP ALA
SEQRES 12  307  ASN ARG ASN TRP ASP ALA GLY PHE GLY LYS ALA GLY ALA
SEQRES 13  307  SER SER SER PRO CYS SER GLU THR TYR HIS GLY LYS TYR
SEQRES 14  307  ALA ASN SER GLU VAL GLU VAL LYS SER ILE VAL ASP PHE
SEQRES 15  307  VAL LYS ASN HIS GLY ASN PHE LYS ALA PHE LEU SER ILE
SEQRES 16  307  HIS SER TYR SER GLN LEU LEU LEU TYR PRO TYR GLY TYR
SEQRES 17  307  THR THR GLN SER ILE PRO ASP LYS THR GLU LEU ASN GLN
SEQRES 18  307  VAL ALA LYS SER ALA VAL ALA ALA LEU LYS SER LEU TYR
SEQRES 19  307  GLY THR SER TYR LYS TYR GLY SER ILE ILE THR THR ILE
SEQRES 20  307  TYR GLN ALA SER GLY GLY SER ILE ASP TRP SER TYR ASN
```

```
SEQRES 21  307  GLN GLY ILE LYS TYR SER PHE THR PHE GLU LEU ARG ASP
SEQRES 22  307  THR GLY ARG TYR GLY PHE LEU LEU PRO ALA SER GLN ILE
SEQRES 23  307  ILE PRO THR ALA GLN GLU THR TRP LEU GLY VAL LEU THR
SEQRES 24  307  ILE MET GLU HIS THR VAL ASN ASN
HELIX    1 H1 THR    14  GLN    28
HELIX    2 H2 GLU    72  GLU .  88
HELIX    3 H3 PRO    94  MET   103
HELIX    4 H4 ASN   112  GLU   122
HELIX    5 H5 GLU   173  GLY   187
HELIX    6 H6 ASP   215  LYS   231
HELIX    7 H7 SER   254  GLY   262
HELIX    8 H8 GLN   285  ASN   306
SHEET    1 S1  LEU   32  LEU   36
SHEET    2 S1  VAL   49  SER   53
SHEET    3 S1  ASP  104  ILE  109
SHEET    4 S1  PRO   60  LEU   66
SHEET    5 S1  LYS  190  HIS  196
SHEET    6 S1  TYR  265  LEU  271
SHEET    7 S1  GLN  200  TYR  204
SHEET    8 S1  LYS  239  GLY  241

COMPND  HEMOGLOBIN (DEOXY)
HEADER  OXYGEN TRANSPORT
SEQRES  1 A  141  VAL LEU SER PRO ALA ASP LYS THR ASN VAL LYS ALA ALA
SEQRES  2 A  141  TRP GLY LYS VAL GLY ALA HIS ALA GLY GLU TYR GLY ALA
SEQRES  3 A  141  GLU ALA LEU GLU ARG MET PHE LEU SER PHE PRO THR THR
SEQRES  4 A  141  LYS THR TYR PHE PRO HIS PHE ASP LEU SER HIS GLY SER
SEQRES  5 A  141  ALA GLN VAL LYS GLY HIS GLY LYS LYS VAL ALA ASP ALA
SEQRES  6 A  141  LEU THR ASN ALA VAL ALA HIS VAL ASP ASP MET PRO ASN
SEQRES  7 A  141  ALA LEU SER ALA LEU SER ASP LEU HIS ALA HIS LYS LEU
SEQRES  8 A  141  ARG VAL ASP PRO VAL ASN PHE LYS LEU LEU SER HIS CYS
SEQRES  9 A  141  LEU LEU VAL THR LEU ALA ALA HIS LEU PRO ALA GLU PHE
SEQRES 10 A  141  THR PRO ALA VAL HIS ALA SER LEU ASP LYS PHE LEU ALA
SEQRES 11 A  141  SER VAL SER THR VAL LEU THR SER LYS TYR ARG
SEQRES  1 B  146  VAL HIS LEU THR PRO GLU GLU LYS SER ALA VAL THR ALA
SEQRES  2 B  146  LEU TRP GLY LYS VAL ASN VAL ASP GLU VAL GLY GLY GLU
SEQRES  3 B  146  ALA LEU GLY ARG LEU LEU VAL VAL TYR PRO TRP THR GLN
SEQRES  4 B  146  ARG PHE PHE GLU SER PHE GLY ASP LEU SER THR PRO ASP
```

```
SEQRES   5 B  146  ALA VAL MET GLY ASN PRO LYS VAL LYS ALA HIS GLY LYS
SEQRES   6 B  146  LYS VAL LEU GLY ALA PHE SER ASP GLY LEU ALA HIS LEU
SEQRES   7 B  146  ASP ASN LEU LYS GLY THR PHE ALA THR LEU SER GLU LEU
SEQRES   8 B  146  HIS CYS ASP LYS LEU HIS VAL ASP PRO GLU ASN PHE ARG
SEQRES   9 B  146  LEU LEU GLY ASN VAL LEU VAL CYS VAL LEU ALA HIS HIS
SEQRES  10 B  146  PHE GLY LYS GLU PHE THR PRO PRO VAL GLN ALA ALA TYR
SEQRES  11 B  146  GLN LYS VAL VAL ALA GLY VAL ALA ASN ALA LEU ALA HIS
SEQRES  12 B  146  LYS TYR HIS
HELIX    1  AA SER     3  GLY    18
HELIX    2  AB HIS    20  SER    35
HELIX    3  AC PHE    36  TYR    42
HELIX    4  AD HIS    50  GLY    51
HELIX    5  AE SER    52  ALA    71
HELIX    6  AF LEU    80  ALA    88
HELIX    7  AG ASP    94  HIS   112
HELIX    8  AH THR   118  SER   138
HELIX    9  BA THR     4  VAL    18
HELIX   10  BB ASN    19  VAL    34
HELIX   11  BC TYR    35  PHE    41
HELIX   12  BD THR    50  GLY    56
HELIX   13  BE ASN    57  ALA    76
HELIX   14  BF PHE    85  CYS    93
HELIX   15  BG ASP    99  HIS   117
HELIX   16  BH THR   123  HIS   143

COMPND    LACTATE DEHYDROGENASE (E.C.1.1.1.27) APO ENZYME M4
HEADER    OXIDOREDUCTASE, CHOH DONOR, NAD ACCEPTR
SEQRES   1  229  ALA THR LEU LYS ASP LYS LEU ILE GLY HIS LEU ALA THR
SEQRES   2  229  SER GLN GLU PRO ARG SER TYR ASN LYS ILE THR VAL VAL
SEQRES   3  229  GLY CYS ASP ALA VAL GLY MET ALA ASP ALA ILE SER VAL
SEQRES   4  229  LEU MET LYS ASP LEU ALA ASP GLU VAL ALA LEU VAL ASP
SEQRES   5  229  VAL MET GLU ASP LYS LEU LYS GLY GLU MET MET ASP LEU
SEQRES   6  229  GLN HIS GLY SER LEU PHE LEU HIS THR ALA LYS ILE VAL
SEQRES   7  229  SER GLY LYS ASP TYR SER VAL SER ALA GLY SER LYS LEU
SEQRES   8  229  VAL VAL ILE THR ALA GLY ALA ARG GLN GLN GLU GLY GLU
SEQRES   9  229  SER ARG LEU ASN LEU VAL GLN ARG ASN VAL ASN ILE PHE
SEQRES  10  229  LYS PHE ILE ILE PRO ASN ILE VAL LYS HIS SER PRO ASP
SEQRES  11  229  CYS ILE ILE LEU VAL VAL SER ASN PRO VAL ASP VAL LEU
```

```
SEQRES  12  229  THR TYR VAL ALA TRP LYS LEU SER GLY LEU PRO MET HIS
SEQRES  13  229  ARG ILE ILE GLY SER GLY CYS ASN LEU ASP SER ALA ARG
SEQRES  14  229  PHE ARG TYR LEU MET GLY GLU ARG LEU GLY VAL HIS SER
SEQRES  15  229  CYS SER GLY VAL GLY TRP VAL ILE GLY GLN HIS GLY ASP
SEQRES  16  229  SER VAL PRO SER VAL TRP SER GLY MET TRP ASN ALA LEU
SEQRES  17  229  LYS GLU LEU HIS PRO GLU LEU GLY THR ASN LYS ASP LYS
SEQRES  18  229  GLN ASP TRP LYS LYS LEU HIS LYS ASP VAL VAL ASP SER
SEQRES  19  229  ALA TYR GLU VAL ILE LYS LEU LYS GLY TYR THR SER TRP
SEQRES  20  229  ALA ILE GLY LEU SER VAL ALA ASP LEU ALA GLU THR ILE
SEQRES  21  229  MET LYS ASN LEU CYS ARG VAL HIS PRO VAL SER THR MET
SEQRES  22  229  VAL LYS ASP PHE TYR GLY ILE LYS ASP ASN VAL PHE LEU
SEQRES  23  229  SER LEU PRO CYS VAL LEU ASN ASP HIS GLY ILE SER ASN
SEQRES  24  229  ILE VAL LYS MET LYS LEU LYS PRO ASN GLU GLU GLN GLN
SEQRES  25  229  LEU GLN LYS SER ALA THR THR LEU TRP ASP ILE GLN LYS
SEQRES  26  229  ASP LEU LYS PHE
HELIX    1  AA THR      2  ILE      8
HELIX    2  AB ASP     29  ASP     43
HELIX    3  AC MET     54  SER     69
HELIX    4  AD GLY    103  ILE    121
HELIX    5  AE PHE    119  SER    128
HELIX    6 A1F VAL    142  SER    151
HELIX    7 A2F CYS    163  GLY    179
HELIX    8 A1G LYS    226  GLU    237
HELIX    9 A2G VAL    238  GLY    243
HELIX   10 A3G ILE    249  LEU    264
HELIX   11  AH PRO    307  ILE    323
SHEET    1 SH1  LYS    76  GLY    80
SHEET    2 SH1  ASP    46  ASP    52
SHEET    3 SH1  ASN    21  GLY    27
SHEET    4 SH1  LYS    90  THR    95
SHEET    5 SH1  ILE   132  VAL   136
SHEET    6 SH1  ARG   157  ILE   159
SHEET    1 SH2  SER   185  SER   185
SHEET    2 SH2  MET   204  ASN   206
SHEET    3 SH2  GLU   210  HIS   212
SHEET    1 SH3  ARG   266  VAL   274
SHEET    2 SH3  VAL   284  LEU   292
SHEET    3 SH3  GLY   296  VAL   301
```

```
COMPND   TROPONIN C
HEADER   CONTRACTILE SYSTEM PROTEIN
SEQRES  1  162  THR SER ALA MET ASP GLN GLN ALA GLU ALA ARG ALA PHE
SEQRES  2  162  LEU SER GLU GLU MET ILE ALA GLU PHE LYS ALA ALA PHE
SEQRES  3  162  ASP MET PHE ASP ALA ASP GLY GLY GLY ASP ILE SER THR
SEQRES  4  162  LYS GLU LEU GLY THR VAL MET ARG MET LEU GLY GLN ASN
SEQRES  5  162  PRO THR LYS GLU GLU LEU ASP ALA ILE ILE GLU GLU VAL
SEQRES  6  162  ASP GLU ASP GLY SER GLY THR ILE ASP PHE GLU GLU PHE
SEQRES  7  162  LEU VAL MET MET VAL ARG GLN MET LYS GLU ASP ALA LYS
SEQRES  8  162  GLY LYS SER GLU GLU GLU LEU ALA ASP CYS PHE ARG ILE
SEQRES  9  162  PHE ASP LYS ASN ALA ASP GLY PHE ILE ASP ILE GLU GLU
SEQRES 10  162  LEU GLY GLU ILE LEU ARG ALA THR GLY GLU HIS VAL THR
SEQRES 11  162  GLU GLU ASP ILE GLU ASP LEU MET LYS ASP SER ASP LYS
SEQRES 12  162  ASN ASN ASP GLY ARG ILE ASP PHE ASP GLU PHE LEU LYS
SEQRES 13  162  MET MET GLU GLY VAL GLN
HELIX    1  NT MET     4  LEU    14
HELIX    2   A GLU    16  PHE    29
HELIX    3   B GLU    41  MET    48
HELIX    4   C LYS    55  ASP    66
HELIX    5 LCH PHE    75  PHE   105
HELIX    6   F GLU   117  THR   125
HELIX    7   G GLU   131  ASP   142
HELIX    8   H GLU   153  GLU   159
```

```
COMPND   TRYPSIN INHIBITOR (CRYSTAL FORM /II$)
HEADER   PROTEINASE INHIBITOR (TRYPSIN)
SEQRES  1  58  ARG PRO ASP PHE CYS LEU GLU PRO PRO TYR THR GLY PRO
SEQRES  2  58  CYS LYS ALA ARG ILE ILE ARG TYR PHE TYR ASN ALA LYS
SEQRES  3  58  ALA GLY LEU CYS GLN THR PHE VAL TYR GLY GLY CYS ARG
SEQRES  4  58  ALA LYS ARG ASN ASN PHE LYS SER ALA GLU ASP CYS MET
SEQRES  5  58  ARG THR CYS GLY GLY ALA
HELIX    1  H1 PRO     2  GLU     7
HELIX    1  H2 SER    47  GLY    56
SHEET    1  S1  LEU   29  TYR    35
SHEET    2  S1  ILE   18  ASN    24
SHEET    3  S1  PHE   45  PHE    45
```

```
COMPND  RIBONUCLEASE A (E.C.3.1.4.22) (JOINT NEUTRON AND X-RAY)
HEADER  HYDROLASE (NUCLEIC ACID,RNA)
SEQRES  1  124  LYS GLU THR ALA ALA ALA LYS PHE GLU ARG GLN HIS MET
SEQRES  2  124  ASP SER SER THR SER ALA ALA SER SER SER ASN TYR CYS
SEQRES  3  124  ASN GLN MET MET LYS SER ARG ASN LEU THR LYS ASP ARG
SEQRES  4  124  CYS LYS PRO VAL ASN THR PHE VAL HIS GLU SER LEU ALA
SEQRES  5  124  ASP VAL GLN ALA VAL CYS SER GLN LYS ASN VAL ALA CYS
SEQRES  6  124  LYS ASN GLY GLN THR ASN CYS TYR GLN SER TYR SER THR
SEQRES  7  124  MET SER ILE THR ASP CYS ARG GLU THR GLY SER SER LYS
SEQRES  8  124  TYR PRO ASN CYS ALA TYR LYS THR THR GLN ALA ASN LYS
SEQRES  9  124  HIS ILE ILE VAL ALA CYS GLU GLY ASN PRO TYR VAL PRO
SEQRES 10  124  VAL HIS PHE ASP ALA SER VAL
HELIX   1  H1 THR     3  MET     13
HELIX   2  H2 ASN    24  ASN     34
HELIX   3  H3 LEU    51  GLN     60
SHEET   1 S1A  LYS   41  HIS     48
SHEET   2 S1A  MET   79  THR .   87
SHEET   3 S1A  ASN   94  LYS    104
SHEET   1 S1B  SER   90  LYS     91
SHEET   1 S2A  LYS   61  ALA     64
SHEET   2 S2A  ASN   71  SER     75
SHEET   3 S2A  HIS  105  ASN    113
SHEET   4 S2A  PRO  114  HIS    119
SHEET   1 S2B  ASP  121  VAL    124

COMPND  LYSOZYME (E.C.3.2.1.17)
HEADER  HYDROLASE (O-GLYCOSYL)
SEQRES  1  129  LYS VAL PHE GLY ARG CYS GLU LEU ALA ALA ALA MET LYS
SEQRES  2  129  ARG HIS GLY LEU ASP ASN TYR ARG GLY TYR SER LEU GLY
SEQRES  3  129  ASN TRP VAL CYS ALA ALA LYS PHE GLU SER ASN PHE ASN
SEQRES  4  129  THR GLN ALA THR ASN ARG ASN THR ASP GLY SER THR ASP
SEQRES  5  129  TYR GLY ILE LEU GLN ILE ASN SER ARG TRP TRP CYS ASN
SEQRES  6  129  ASP GLY ARG THR PRO GLY SER ARG ASN LEU CYS ASN ILE
SEQRES  7  129  PRO CYS SER ALA LEU LEU SER SER ASP ILE THR ALA SER
SEQRES  8  129  VAL ASN CYS ALA LYS LYS ILE VAL SER ASP GLY ASP GLY
SEQRES  9  129  MET ASN ALA TRP VAL ALA TRP ARG ASN ARG CYS LYS GLY
SEQRES 10  129  THR ASP VAL GLN ALA TRP ILE ARG GLY CYS ARG LEU
```

HELIX	1	A	ARG	5	HIS	15
HELIX	2	B	LEU	25	GLU	35
HELIX	3	C	CYS	80	LEU	84
HELIX	4	D	THR	89	LYS	96
SHEET	1	S1	LYS	1	PHE	3
SHEET	2	S1	PHE	38	THR	40
SHEET	1	S2	ALA	42	ASN	46
SHEET	2	S2	SER	50	GLY	54
SHEET	3	S2	GLN	57	SER	60

APPENDIX.C.II.1

```
Human Spl
Q H I C H I   Q G C G K V Y G K T S H L R A H L R     W H T G
P F M C T W   S Y C G K R F T R S D E L Q R H K R     T H T G
K F A C P E       C P K R F M R S D H L S K H I K     T H Q N

Drosophila Serendipity
E I P C H I       C G E M F S S Q E V L E R H I K A D T C Q K
Q A T C N V       C G L K V K D D E V L D L H M N     L H E G
E L E C R Y       C D K K F S H K R N V L R H M E     V H W D
K Y Q C D K       C G E R F S L S W L M Y N H L M     R H D A
A L I C E V       C H Q Q F K T K R T Y L H H L R     T H Q T
  Y P C P D       C E K S F V D K Y T L K V H K R     V H Q P

Drosophila Serendipity
K Q E C T T       C G K V Y N S W Y Q L Q K H L S   E E H S K
N H I C P I       C G V I R R D E E Y L E L H M N     L H E G
E K Q C R Y       C P K S F S R P V N T L R H M R     S H W D
K Y Q C E K       C G L R F S Q D N L L Y N H R L     R H E A
P I I C S I       C N V S F K S R K T F N H H T L     I H K E
  H Y C S V       C P K S F T E R Y T L K M H M K     T H E G
S G F C L I       C N T T F E N K K E L E H H L Q     F H A D

Drosophila Kruppel
S F T C K I       C S R S F G Y K H V L Q N H E R     T H T G
P F E C P E       C D K R F T R D H H L K T H M R     L H T G
P Y H C S H       C D R Q F V Q V A N L R R H L R     V H T G
P Y T C E I       C D G K F S D S N Q L K S H M L     V H T G
P F E C E R       C H M K F R R R H H L M N H K         C G I

Drosophila Snail
R F K C D E       C Q K M Y S T S M G L S K H R Q     F H C P
T H S C E E       C G K L Y T T I G A L K M H I R       H T L
P C K C P I       C G K A F S R P W L L Q G H I R     T H T G
P F Q C P D       C P R S F A D R S N L R A H Q Q     T H V D
K Y A C Q V       C H K S F S R M S L L N K H S S   S N C T I
```

Xenopus Xfin

S H H C P H	C K K S F V Q R S V F L K H Q R	T H T G
P Y Q C V E	C Q K K F T E R S A L V N H Q R	T H T G
P Y T C L D	C Q K T F N Q R S A L T K H R R	T H T G
P Y R C S V	C S K S F I Q N S D L V K H L R	T H T G
P Y E C P L	C V K R F A E S S A L M K H K R	T H S T
P F R C S C	C S R S F T H N S D L T A H M R	K H T E
P Y S C S K	C R K T F K R W K S F L N H Q Q	T H S R
P Y L C S H	C N K G F I Q N S D L V K H F R	T H T G
P Y Q C A E	C H K G F I Q K S D L V K H L R	T H T G
P F K C S H	C D K K F T E R S A L A K H Q R	T H T G
P Y K C S D	C G K P F T Q R S N L I L H Q R	I H T G
P Y K C T L	C D R T F I Q N S D L V K H Q K	V H A N
P H K C S K	C D L T F S H W S T F M K H S K	L H S G
K F Q C A E	C K K G F T Q K S D L V K H I R	V H T G
P F K C L L	C K K S F S Q N S D L H K H W R	I H T G
P F P C Y T	C D K S F T E R S A L I K H H R	T H T G
P H K C S V	C Q K G F I Q K S A L T K H S R	T H T G
P Y P C T Q	C G K S F I Q N S D L V K H Q R	I H T G
P Y H T C E	C N K R F T E G S S L V K H R R	T H S G
P Y R C P Q	C E K T F I Q S S D L V K H L V	V H N G
P Y P C T E	C G K V F H Q R P A L L K H L R	T H K T
R Y P C N E	C S K E F F Q T S D L V K H L R	T H T G
P Y H C P E	C N K G F I Q N S D L V K H Q R	T H T G
P Y T C S Q	C D K G F I Q R S A L T K H M R	T H T G
P Y K C E Q	C Q C K F I Q N S D L V K H Q R	I H T G
P Y H C P D	C D K R F T E H S S L I K H Q R	I H S R
P Y P C G V	C G K S F S Q S S N L L K H L K	C H S E
S F K C N D	C G K C F A H R S V L I K H V R	I H T G
P Y K C S Q	C T R S F I Q K S D L V K H Y R	T H T G
P Y K C G L	C E R S F V E K S A L S R H Q R	V H K N
R Y S C S E	C G K C F T H R S V F L K H W R	M H T G
P Y T C K E	C G K S F S Q S S A L V K H V R	I H T G
P Y A C S T	C G K S F I Q K S D L A K H Q R	I H T G
P Y T C T V	C G K K F I D R S S V V K H S R	T H T G

P Y K C N E   C T K G F V Q K S D L V K H M R   T H T G
P Y G C N C   C D R S F S T H S A S V R H Q R   M C N T

Drosophila Terminus

D L H C R R   C R T Q F S R R S K L H I H Q K   L R C G Q

Yeast SW15

T F E C L F   P G C T K T F K R R Y N I R S H I Q   T H L E
P Y S C D H   P G C D K A F V R N H D L I R H K K   S H Q E
Y A C P   C G K K F N R E D A L V V H R S R M I C S G

Xenopus Transcription Factor IIIA

R Y I C S F   A D C G A A Y N K N W K L Q A H L C   K H T G
P F P C K E   E G C E K G F T S L H H L T R H S L   T H T G
N F T C D S   D G C D L R F T T K A N M K K H F N   R F H N I
V Y V C H F   E N C G K A F K K H N Q L K V H Q F   S H T Q
P Y E C P H   E G C D K R F S L P S R L K R H E K   V H A G
Y P C K K D D S C S F V G K T W T L Y L K H V A   E C H Q D
A V C D V   C N R K F R H K D Y L R D H Q K   T H E K
V Y L C P R   D G C D R S Y T T A F N L R S H I Q   S F H E E
P F V C E H   A G C G K C F A M K K S L E R H S V   V H D P

Yeast ADRI

S F V C E V   C T R A F A R Q E H L K R H Y R   S H T N
P Y P C G L   C N R C F T R R D L L I R H A Q   K I H S G

Drosophila Krh

T Y R C S E   C Q R E F E L L A G L K K H L K   T H R T
K Y Q C D I   C G Q K F V Q K I N L T H H A R   I H S S
P Y E C P E   C Q K R F Q E R S H L Q R H Q K   Y H A Q
S Y R C E K   C G K M Y K T E R C L K V H N L   V H L E
P F A C T V   C D K S F I S N S K L K Q H S N   I H T G
P F K C N Y   C P R D F T N F P N W L K H T R   R R H K V

Mouse Mkr1

P F V C N Y   C D K T F S F K S L L V S H K R   I H T G

P Y E C D V   C Q K T F S H K A N L I K H Q R   I H T G
P F E C P E   C G K A F T H Q S N L I V H Q R   A H M E
P Y G C S E   C G K A F T H Q S N L I V H Q R   I H T G
P Y E C N E   C A K T F F K K S N L I I H Q K   I H T G
R Y E C S E   C G K S F I Q N S Q L I I H R R   T H T G
P Y E C T E   C G K T F S Q R S T L R L H L R   I H T G

Mouse Mkr2

V Y G C D E   C G K T F R Q S S S L L K H Q R   I H T G
P Y T C N V   C D K H F I E R S S L T V H Q R   T H T G
P Y K C H E   C G K A F S Q S M N L T V H Q R   T H T G
P Y Q C K E   C G K A F R K N S S L I Q H E R   I H T G
P Y K C H D   C E K A F S K N S S L T Q H R R   I H T G
P Y E C M I   C G K H F T G R S S L T V H Q V   I H T G
P Y E C T E   C G K A F S Q S A Y L I E H R R   I H T G
P Y E C D Q   C G K A F I K N S S L I V H Q R   I H T G
P Y Q C N E   C G K P F S R S T N L T R H Q R   T H T

Drosophila Hunchback

N Y K C K T   C G V V A I T K V D F W A H T R   T H M K
I L Q C P K   C P F V T E F K H H L E Y H I R   K H K N
P F Q C D K   C S Y T C V N K S M L N S H R K   S H S S
Q Y R C A D   C D Y A T K Y C H S F K L H L R H Y G H K P
I Y E C K Y   C D I F F K D A V L Y T I H M G   Y H S C
V F K C N M   C G E K C D G P V G L F V H M A R N A H S

Trypanosome TRS-1

P T K C T E   C D A T Y Q C R S S A V T H M V   N K H G F
V L H C T I   C A S K F A V P G R L L H H L R   T I H G I
P F Q C D L   C E A S F G T H S S L S L H K K   L K H K S
E V Q C G V   C Q K V L S C R D S L I R H C K   A F H K G
M L V C P T   C G R Q C A S K T G L T L H Q K   K M H G M

Mouse NGFI-A

P Y A C P V   E S C D R R F S R S D E L T R H I R   I H T G
P F Q C R I   C M R N F S R S D H L T T H I R   T H T G

P F A C D I	C G R K F A R S D E R K R H T K	I H L R

Human ZFY

V Y P C M I	C G K K F K S R G F L K R H M K	N H P E
K Y H C T D	C D Y T T N K K I S L H N H L E	S H K L
A I E C D E	C G K H F S H A G A L F T H K M	V H K E
M H K C K F	C E Y E T A E Q G L L N R H L L	A V H S K
P H I C V E	C G K G F R H P S E L R K H M R	I H T G
P Y Q C Q Y	C E Y R S A D S S N L K T H I K	T K H S K
P F K C D I	C L L T F S D T K E V Q Q H T L	V H Q E
T H Q C L H	C D H K S S N S S D L K R H V I	S V H T K
P H K C E M	C E K G F H R P S E L K K H V A	V H K G
M H Q C R H	C D F K I A D P F V L S R H I L	S V H T K
P F R C K R	C R K G F R Q Q N E L K K H M K	T H S G
V Y Q C E Y	C E Y S T T D A S G F K R H V I	S I H T K
P H R C E Y	C K K G F R R P S E K N Q H I M	R H H K

Xenopus p43 5S RNA Binding Protein

L L R C P A	A G C K A F Y R K E G K L Q D H M A	G H S E
P W K C G I	K D C D K V F A R K R Q I L K H V K	R H L A
K L S C P T	A G C K M T F S T K K S L S R H K L	Y K H G E
P L K C F V	P G C K R S F R K K R A L R R H L S	V H S N
L S V C D V	P G C S W K S S S V A K L V A H Q K	R H R G
Y R C S Y	E G C Q T V S P T W T A L Q T H V K	K H P L
L Q C A A	C K K P F K K A S A L R R H K A	T H A K
Q L P C P R	Q D C D K T F S S V F N L T H H V A	R K L H L C
T H R C P H	S G C T R S F A M R E S L L R H L V	V H D P

APPENDIX.C.III.1

1. Antamanide

   [V-P-P-A-F-F-P-P-F-]

2. Anaphylatoxin

   A-S-H-L-G-L-A-R

3. Bacitracin-F

   L-F-I-[-K-Orn-I-F-H-D]-D(NH_2)

4. Bacillomycin L

   [β Aminoacid-D-Y*-D*-S-E-S*-T-]

5. Bacitracin A

   $$NH_2\text{-}\underbrace{\text{I-C}}_{S}\text{-L-E}^{*}\text{-L-E}^{*}\text{-I-[K-Orn-I-F}^{*}\text{-H-D}^{dl}\text{]-D}^{dl}\text{-COOH}$$

6. Cycloamanide-I

   [P-S-F-F-F-P-I-]

7. Cycloamanide-II

   [-P-M-L-G-F-L-V-L-]

8. Cycloamanide-III

   [-P-M-L-G-F-L-P-L-]

9. Bottromycin A_2

   [Subs.-G-P-V-]

10. BE-4

    [A-L-D-A-Subs.-IsoQ]

11. Bacillus Subtilis C-756 metabolite

    $$R(CH_2)_6[CH\text{-}CH_2\text{-}\overset{\overset{\displaystyle O}{\|}}{C}\text{-E-L-L-V-D-L-L-}]$$

12. Cyclolinopeptide

[-L-I-I-L-V-P-P-F-F-]

13. Desthiomalformin

NH_2-I-A*-A*-V-L*-OH

14. Deoxybouvardin

[-Y-A-A-Y-A-Y-]

15. Didemnin-A

L-[T-Subs.-L-P-W-]

16. Evolidine

[-S-F-L-P-V-N-L-]

17. Enkephalin

```
Y-S-G-F-S-OH
  |     |
 -Si-O-Si-
```

18. Echinocandin-D

[p-hydrophenyl T-P-T-Subs.-P-T-]

19. Fungisporin

[-F*-F-V*-V-F*-F-V*-V-]

20. Gramicidin-A

HCO-V-G-A-L*-A-V*-V-V*-W-L*-W-L*-W-L*-W-NHC_2H_4OH

21. Gramicidin-SA

[V-Orn-L-F*-P-V-Orn-L-F*-P-]

22. Gramicidin-J_1

[V-Orn-F*-L-*-F-P-Orn-]

23. Gramicidin-J_2

[V-Orn-L*-F-*-P-Orn-]

24. Griselimycin

MeCO-meV-P-[meT-L-P-L-meV-P-meL*-G-]

25. Gratisin

[L-F-P-Y-V-Orn-L-F-P-Y-V-Orn]

26. Leupeptin

Me-C0-L-L-R

27. Longicatenamycin

[L-I*-K*-βhydroxyE-5chloroW*-G-]

28. Lophyrotomin

ph-A*-F*-V-I-D*-D-E*-Q-OH

29. Malformin A

C*-V-C*-L*-I

30. Malformin C

C*-C*-V-L*-L*

31. Mulndocandin

[P-$Subs_1$-T-P-$Subs_2$-S-]

32. Mycobacillin

[-A-D*-P-D*-E*-Y-D-Y-S-D*-L-E*-D*-]

33. Mycosubtilin

[-N-Q-P-Y*-N*-N*-S-D-βAmines-]

34. Norsufactin

Me$(CH_2)_{10}$-[CH-CH_2-CO-E-L-L-V-D-L-L-]

35. Polymyxin E

$$\begin{array}{l} \qquad\qquad\qquad\qquad\quad \text{NH-NH}_2 \\ \qquad\qquad\qquad\qquad\quad | \\ \text{Pelargonyl-K-T-K-K-T-K-L*-K} \end{array}$$

36. Polymyxin M

$$\begin{array}{l} \qquad\qquad\qquad\qquad\quad \text{NH-NH}_2 \\ \qquad\qquad\qquad\qquad\quad | \\ \text{Pelargonyl-K-T-K-K-T-K-T*-K} \end{array}$$

37. Peptidoglycan

A-γE*-K-A*-A*

38. Phalloin

[-A-W-L-A-A-C-P-]

39. Retroantamanide

[-F-F-P-P-F-F-A-P-P-V-]

40. Retrogramicidin S

[-F*-L-Orn-V-P-F*-L-Orn-V-P-]

41. Tyrocidine A

[-V-Orn-L-F*-P-F-F*-N-Q-Y-]

42. Tyrocidine B

[-V-Orn-L-F*-P-Y-F*-N-Q-Y-]

43. Tyrocidine C

[-V-Orn-L-F*-P-Y-W*-N-Q-Y-]

44. Tyrocidine E

[-V-Orn-L-F*-P-F-F*-N-Q-F-]

45. Tuberactinomycin

[-H-E-C-F*-G-H-E-C-F*-G-]

46. Terlipressin

G-G-G-C-Y-F-Q-N-C-P-K-G-NH_2

47. TL-119

AcF-L-F-[T-V-A-$\alpha\beta$hydroxyT-]

48. Rhozonin A

[-I-V-V-βfurylA-L-A-βfurylA-]

49. Rhozonin B

[-V-V-V-βfurylA-L-A-βfurylA-]

50. Viscunamide

[-L*-I-L*-I-L*-]

51. HC-Toxin

Aoe-P*-A-A*-]

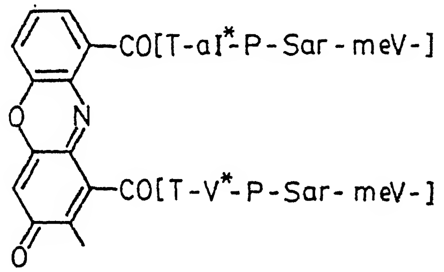
52. Viscosin

L-G-S-V-T

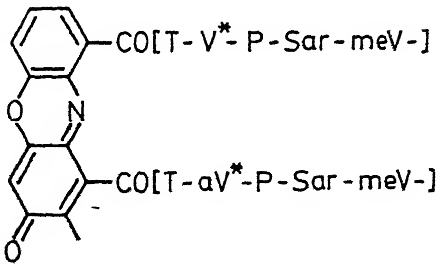
53. Gramicidin S

[-V-Orn-V-F*-P-V-Orn-V-F*-P-]

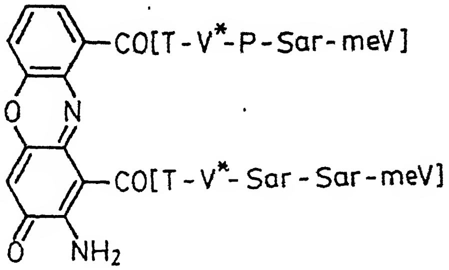
54. Actinomycin-C_2



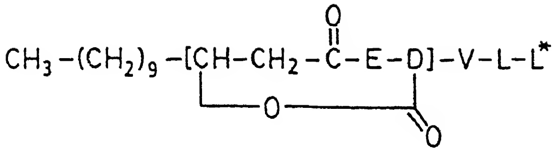
55. Aniso-Actinomycin D



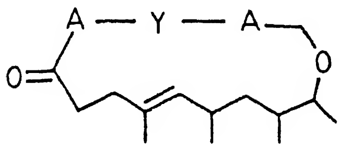
56. Actinomycin.X



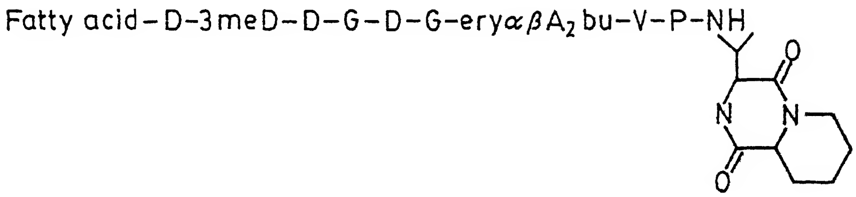
57. Esperin



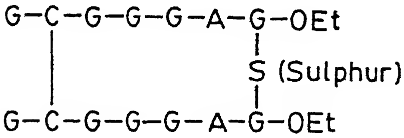
58. Geodiamolides



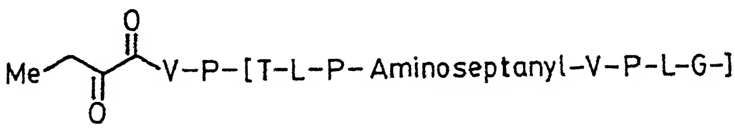
59. Glumamycin



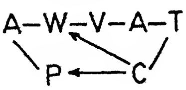
60. Lanthionine



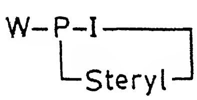
61. Mycoplanecin A



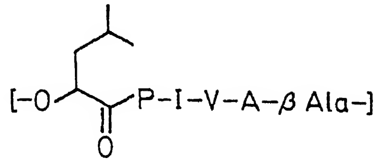
62. Norphalloin



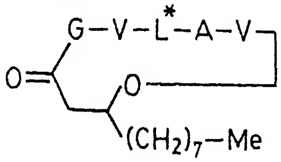
63. Nummularia



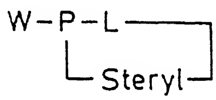
64. Protodestruxin



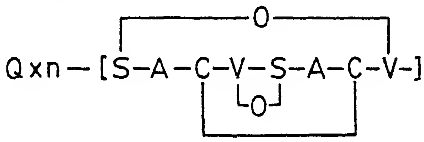
65. Isarin



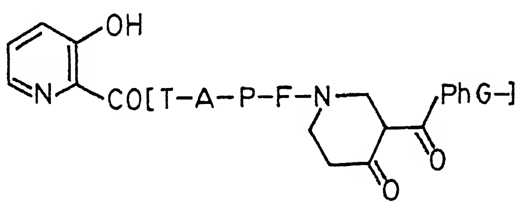
66. Sativanine E



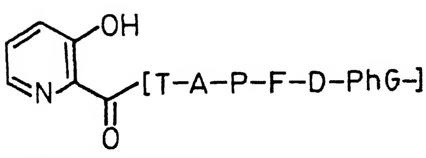
67. Triostin A



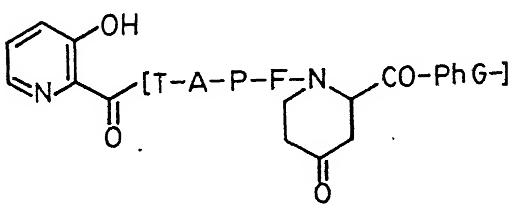
68. Vernamycin B



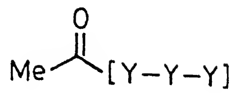
69. Dorcidin



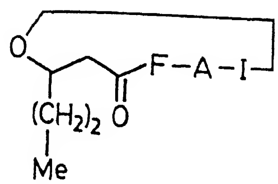
70. Ostreogrycin B



71. K-13



72. Beavellide



73. Dextruxin